Ana Elisa Soares Carneiro

Formação em Aveiro: análise estatística num estágio na Multidados

Ana Elisa Soares Carneiro

Estudo sobre Formação – Análise de Dados Categorizados num Estágio na Multidados

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Matemática e Aplicações, área de especialização Matemática Empresarial e Tecnológica, realizada sob a orientação científica da Doutora Cláudia Margarida Pedrosa Neves, Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro.

o júri / the jury

presidente / president	**Professora Doutora Isabel Maria Pereira Simões** Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro
vogais / examiners committee	**Professor Doutor Luis Miguel Domingues Fernandes Ferreira** Professor Auxiliar da Universidade de Aveiro
vogais / examiners committee	**Doutora Filomena Frazão** Técnica da Multidados (co-orientadora)
	Professora Doutora Cláudia Margarida Pedrosa Neves Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro (orientadora)

agradecimentos / acknowledgements

À minha orientadora, Professora Doutora Cláudia Neves, pela sua orientação científica e disponibilidade.
À minha família, pelo carinho e compreensão demonstrados ao longo de todo este percurso.
Aos amigos pelo apoio e pelas palavras de incentivo e encorajamento.

Palavras-chave Amostragem, Estimação, Formação, Modelos Log-Lineares, Proporção, Tabelas de Contingência

Resumo Este trabalho tem como objectivo principal o de fornecer um estudo detalhado sobre a formação de variada natureza (não necessariamente formação profissional) que se procura e se realiza no Distrito de Aveiro, tanto a nível pessoal como em empresas. A análise estatística para o efeito incide sobre a estimação de proporções em dados categorizados. Também estão presentes procedimentos adequados de análise da relação entre as variáveis escolhidas para caracterizar a formação. O estudo é alicerçado em informação recolhida através de questionários aplicados via telefone, tomando em consideração a uma medida de incidência regional.

Keywords Contingency Tables, Estimation, Log-Linear Models, Proportion, Sampling, Training

Abstract The present work aims at providing a thorough assessment of the offer and demand for training of various nature (not only professional training) in Aveiro, at two levels, personal and corporate. To this effect, the adopted statistical procedures rely heavily on statistics for proportions pertaining to categorical data. Statistical procedures for detecting association between variables are also addressed. This survey consists of data records obtained by means of an enquiry applied via telephone, by taking a specific incidence measure into account.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A formação ao longo da vida tornou-se essencial nos dias de hoje pelas mais variadas razões, desde a evolução profissional, valorização pessoal, definida pela situação económica ou até, mais simplesmente, por interesse e gosto pessoal. A realização de formação permite adquirir novos conhecimentos em áreas desconhecidas pelo indíviduo até então, assim como o aprofundamento de conceitos anteriormente adquiridos. As empresas investem cada vez mais em cursos de formação para os trabalhadores, para que possam desempenhar a sua função de forma mais eficaz, e estejam preparados para enfrentar desafios no meio empresarial, que se tornou um meio bastante competitivo. Paralelamente à formação nas componente técnicas, também se verifica uma aposta cada vez maior em formaçao na área comportamental.

Tendo em vista esta crescente necessidade e procura de formação, foi realizado um estudo de investigação do tema, no âmbito do estágio curricular do Mestrado em Matemática e Aplicações, Ramo Matemática Empresarial e Tecnológica, suscitado pela empresa MultiDados® que orientou o estágio com a sua vasta experiência em estudos de mercado e tratamento estatístico de dados.

O objectivo principal deste estudo é analisar o nível e tipo de formação não académica mais premente na população do distrito de Aveiro. Esta população, de agora em diante referida como "Particulares", é composta pelos residentes no distrito de Aveiro com idade igual ou superior a 18 anos. Também se pensou ser importante obter o mesmo tipo de informação em relação às empresas deste mesmo distrito, dado que é natural existir alguma relação entre formação particular e profissional, sendo esta última a população "Empresas".

Assim pretende-se obter respostas a questões do tipo que se segue:

No caso dos particulares,

- Qual a proporção de pessoas do distrito de Aveiro que fizeram algum tipo de formação?
- Em que áreas se fazem mais formações?
- Quais os motivos da frequência de formações?

- Que condições são mais relevantes em relação ao tipo de formação?
- As formações melhoraram algum aspecto pessoal ou profissional?
- Quais os motivos da não frequência de formações?
- Que factores poderiam contribuir para o aumento da frequência de formações?

Sobre as empresas,

- Qual a proporção (ou número) de empresas do distrito de Aveiro que disponibilizam formação aos colaboradores?
- Em que áreas realizam formação?
- Quais os motivos de realizarem formação?
- Qual a relação áreas de formação realizada/área de actividade da empresa?
- A formação realizada ajudou de algum modo a produtividae dos colaboradores?
- Que condições são mais relevantes para cada tipo de formação?

Como preparação para a recolha dos dados foi necessário elaborar um guião sob a orientação da Multidados, reportando-se ao objetivo do estudo, e de forma a providenciar um inquérito que fornecesse informação útil para os aspectos acima referidos em forma de questão. Como não seria adequado fazer exactamente as mesmas perguntas para particulares e empresas optou-se por realizar dois inquéritos distintos, um direcionado para os particulares e o outro para as empresas.

Posteriormente surge a necessidade do conhecimento da população a estudar e consequentemente da amostra que será usada. Os dados sobre o número de pessoas e empresas foram retirados, respectivamente, do site do INE e de um site com o directório de empresas de Portugal, encontrando-se em anexo.

Esta dissertação está dividida em 5 capítulos. Neste capítulo foi apresentada a motivação deste trabalho e o seu enquadramento no âmbito do estágio curricular realizado, assim como foram enunciados os objectivos gerais que se pretendem atingir com o estudo do tema.

No capítulo 2 aborda-se a amostragem estatística, que é uma parte essencial em qualquer estudo estatístico. São descritos de sumariamente os planos de amostragem mais comuns, e em detalhe o plano que foi utilizado neste trabalho, demonstrando-se as propriedades mais relevantes dos estimadores usados, nomeadamente a consistência e a normalidade assintótica. Assim, foi possível especificar o erro máximo associado à estimação através de intervalos de confiança, bem como, e sempre que adequado, proceder a testes de hipóteses sobre os parâmetros em causa. Refira-se aqui, que a análise efetuada neste estudo comporta essencialmente técnicas de dados categorizados e portanto, a medida de probabilidade (com leitura em termos de proporção) assume aqui especial relevância.

No caso multivariado, para averiguar acerca da natureza da relação entre algumas variáveis, introduz-se no capítulo 3 tabelas de contingência. Primeiramente são descritas tabelas de contingência para dois fatores, das quais se obterão conclusões sobre o tipo de associação entre algumas das variáveis em estudo, através de testes de independência e de testes de homogeneidade do Qui-Quadrado. Depois da análise de dois factores, refere-se também tabelas com três fatores, assim como os tipos de independência existentes neste caso.

Ainda sobre relações entre três variáveis, no capítulo 4, são apresentados modelos lineares generalizados, em particular modelos log-lineares e modelos logísticos. Os primeiros são uma alternativa às tabelas de contingência no sentido em que também se investigam a associação entre duas, três ou mais variáveis, mas através de um método diferente.

No capítulo 5 encontram-se as concluões finais desta dissertação.

Em cada um dos capítulos, o tópico discutido será sempre seguido da aplicação ao caso em estudo. Todos os cálculos e funções necessários neste trabalho foram realizados através do *software* estatístico R. [16]

Capítulo 2

Amostragem

2.1 Introdução

Introduz-se neste capítulo o conceito de Amostragem. Existe uma necessidade constante de se obter informação de qualidade e conhecer características relevantes acerca de uma certa população, nas mais variadas áreas do conhecimento. No entanto, por variada ordem de razão, nem sempre é possível estudar de modo compreensivo todos os elementos da população. Por exemplo, a população pode ter dimensão infinita; mesmo no caso de populações finitas, o tempo e recursos disponíveis podem ser limitados, o custo de inquirir toda a população pode ser excessivo, inacessibilidade a alguns elementos da população. Por isso se recorre à Amostragem da população.

Neste enquadramento, definem-se dois conceitos bastante importantes no delineamento de um estudo estatístico: População e Amostra. Designa-se por População o conjunto de indivíduos ou elementos sobre os quais vai incidir o estudo a realizar, e que expressam uma ou mais caraterísticas de interesse para o estudo. Esta pode ser finita ou infinita, e deve ser definida num determinado espaço e tempo. Amostra é um subconjunto da população, através do qual se estudam as referidas caraterísticas. São os elementos da amostra que vão ser entrevistados, estudados ou medidos.

A amostragem consiste, assim, no processo de seleção de uma amostra representativa da população alvo. A partir da amostra, extrai-se a informação necessária, para depois se poder extrapolar as conclusões para a população de onde foi recolhida. Desta forma, será possível obter informação do todo a partir de uma sua parte, isto é, fazer inferência estatística. Geralmente, o objetivo é a estimação de parâmetros da população como, por exemplo, a média, a variância e a proporção da variável de interesse.

Uma amostra diz-se representativa da população se não existirem razões para duvidar que uma dada característica possa diferir da amostra para a população. Se a amostra não representar correctamente a variabilidade e a diversidade que existe na população, diz-se enviesada, e a sua utilização pode dar origem a conclusões erradas comprometendo a fiabilidade dos resultados. A credibilidade das estimativas conseguidas depende do quão bem a amostra foi escolhida e do quão bem as caraterísticas foram obtidas.

2.2 Planos de Amostragem

Para selecionar os elementos a integrar a amostra utilizam-se planos de amostragem, claramente definidos. Tais planos podem ser agrupados em duas grandes classes, com base no modo como a amostra foi selecionada: Amostragem Probabilística e Amostragem Não-Probabilística.

Em Amostragem Probabilística, cada elemento da população tem uma probabilidade conhecida (não nula) de ser selecionado para pertencer à amostra. A aleatoriedade destes métodos é condição necessária para a amostra ser representativa da população, pois qualquer afastamento entra a população e a amostra resulta do acaso. Esta aleatoriedade também evita o enviesamento da amostra e possibilita a aplicação das leis da probabilidade. Uma das vantagens dos métodos probabilísticos é que as estimativas obtidas podem ser extrapoladas para a população e a precisão é mensurável. A distribuição amostral pode ser determinada (pelo menos assintoticamente). Assim é possível calcular o seu desvio-padrão que indica a precisão do processo adotado, medindo o erro amostral decorrente da utilização de apenas uma parte da população.

Nos métodos Não Probabilísticos, o processo de seleção da amostra não é aleatório, pois os elementos que farão parte da amostra são de certa forma escolhidos pelo autor do estudo, com base em critérios de conveniência e julgamento. As vantagens deste tipo de amostragem em relação aos métodos Probabilísticos é que não requer tanto tempo de planeamento e execução, sendo também por isso menos dispendiosa. No entanto, não é possível generalizar os resultados obtidos da amostra para a população, já que não se conhece a probabilidade de um elemento ser incluida na amostra, e assim não existe nenhum modo de determinar a validade dos resultados obtidos. Por este motivo, neste trabalho será utilizado um procedimento de Amostragem Probabilística.

De entre os métodos probabilísticos, existem quatro planos de amostragem mais gerais de seleção de uma amostra de dimensao n, de entre os N elementos que compõem a população $\mathscr{P}$. Descrevem-se de seguida estes planos de modo sucinto:

- **Amostragem aleatória simples (com reposição ou sem reposição)**
  Neste tipo de amostragem, sem reposição, a amostra é obtida selecionando aleatoriamente n elementos distintos de entre os N elementos da população, de tal forma que cada conjunto possível de n elementos tem a mesma probabilidade de ser a amostra selecionada. Assim, a probabilidade de um elemento estar na amostra é de n/N e a probabilidade de uma amostra ser selecionada de entre as amostras possíveis é $1/\binom{N}{n}$. Geralmente, é atribuido um número de 1 a N a cada elemento da população, e através de algum processo aleatório, como programas geradores de números aleatórios ou tabelas de números aleatórios, obtêm-se um conjunto de n números. Os elementos da população correspondentes a estes números irão constituir a amostra.
  Caso exista reposição, cada elemento pode entrar na amostra mais do que uma vez, tendo cada um deles probabilidade $1/N$ de pertencer à amostra. As extracções são independentes e neste caso, a probabilidade de uma amostra ser selecionada de entre as amostras possíveis é $1/N^n$.

O plano de amostragem aleatória simples sem reposição poderia ser, por exemplo, aplicado à situação em que se pretende conhecer o número de empresas, de entre uma população de $N = 350$ empresas, que obtiveram lucros no último ano. Poder-se-ia aplicar este plano para obter uma amostra representativa da população com dimensão $n = 100$ empresas.

- **Amostragem aleatória sistemática**
  Suponha-se que as N unidades da população são numeradas de 1 a N e ordenadas por algum critério. Determina-se primeiro o intervalo de amostragem K que corresponde à parte inteira do quociente N/n. Uma amostra aleatória sistemática de dimensão n é obtida selecionando aleatoriamente um elemento de entre os K primeiros elementos, e daqui, adicionando-se à amostra todos os k-ésimos a partir do último elemento que entrou na amostra. Este método é geralmente utilizado quando se tem acesso a uma lista completa da população.
  É por exemplo o caso dos registos hospitalares. Suponhamos que determinado hospital detém $N = 1000$ fichas de pacientes ordenados por ordem alfabética, e pretende-se obter uma amostra de 500 pacientes. O intervalo de amostragem é $10000/500 = 20$. Escolhe-se aleatoriamente um número de 1 a 20 e a partir desse elemento, os seguintes são selecionados de 20 em 20.

- **Amostragem estratificada**
  Em amostragem estratificada os elementos da população são divididos em grupos ou estratos que partilham alguma característica. Os estratos são mutuamente exclusivos e exaustivos e devem ser o mais homogéneos possível relativamente aos elementos que os constituem. De seguida, recolhe-se uma amostra de cada um dos estratos por amostragem aleatória simples. A amostra pretendida será o conjunto das amostras recolhidas de cada um dos estratos.
  Suponhamos, por exemplo, que numa dada escola se pretende determinar a proporção de alunos de 10 anos com cáries. É aceitável supor que a incidência de cárie dentária dependa do nível socioeconómico da criança, logo considera-se à partida a população dividida em estratos, tantos quantos o número de níveis socioeconómicos existentes.

- **Amostragem por grupos (clusters)**
  Aqui também se divide a população em grupos homogéneos, todos com as mesmas unidades elementares da população. No entanto, a amostra será formada por todos os elementos de alguns destes grupos, que são selecionados mais uma vez por amostragem aleatória simples. Normalmente, este método de amostragem é aplicado quando não se tem uma lista completa da população.
  Para determinar a propoção de aprovações dos alunos de uma escola, tendo em posse a listagem das suas turmas, escolhe-se aleatoriamente uma amostra dessas turmas e entrevistam-se todos os alunos pertencentes às turmas selecionadas. Este é um exemplo em que o plano de amostragem por grupos é apropriado.

Tendo tudo em consideração, a mensagem fundamental é a de que o plano de amostragem deve ser escolhido tendo em conta as caraterísticas da população e os objetivos do estudo, para que se possa obter uma precisão e eficiência máxima nos estimadores adoptados com o mínimo de custos, tempo e recursos dispendidos.

A amostragem estratificada é mais eficiente do que os métodos de amostragem simples ou sistemática, pois usa informação existente sobre a população fornecendo assim resultados com menor probabilidade de erro associada. Também é mais económico em termos de tempo e dinheiro. Relativamente a outros métodos de amostragem, a amostragem aleatória simples não seria adequada pois requer que todos os elementos da população apresentem carterísticas homogéneas entre si, para que se possam selecionar aleatoriamente elementos para a amostra. A amostragem aleatória sistemática também não seria apropriada dado que os elementos da população não se encontram ordenados. A amostragem aleatória por conglomerados também não é indicada pois a população não pode ser dividida em grupos heterogéneos, para depois serem selecionados alguns, de modo aleatório.

2.3 Amostragem Estratificada

Para este estudo optou-se pelo plano de amostragem estratificada, devido ao facto de a população que se vai estudar, população residente no Distrito de Aveiro, se apresentar dividida por municípios, todos com diferentes dimensões, tal como as empresas estão divididas por setores de atividade económica também de dimensão distinta entre si. Consequentemente haveria uma grande chance de a amostra conter mais elementos de um determinado município, ou menos de outro, devido às diferentes dimensões de cada município, e deste modo é assegurado que toda a amostra está "espalhada" por toda a área em estudo. Outro dos motivos é porque se pode supôr, intuitivamente, que partes da população apresentam comportamentos substancialmente diferentes relativamente à variável de interesse, o que acontece com a população em estudo.

Neste tipo de amostragem, a população é dividida em conjuntos, denominados estratos, com base numa dada variável que todos os elementos tenham em comum (variável de estratificação). Estes devem ser definidos de tal modo que dentro de cada estrato exista o máximo de homogeneidade possível relativamente à variável em estudo, mas que sejam heterogéneos entre si. Ou seja, a amostragem estratificada é adequada quando existe grande variabilidade entre os conjuntos ou regiões (estratos) e reduzida variabilidade dentro deles. Estando a população dividida em estratos, fazemos incidir um plano de amostragem aleatória simples dentro de cada estrato. A amostra completa com a qual se vai trabalhar é obtida através da reunião das amostras de cada um dos estratos. Neste caso designa-se o método como amostragem aleatória estratificada (a.a.e.).

Um passo importante consiste em decidir qual a forma mais adequada de afetar as unidades amostrais necessárias ao longo dos vários estratos. Como as dimensões dos estratos definidos neste estudo são conhecidas e diferem todas entre si, deve ser aplicada afetação proporcional, em que o número de elementos de cada estrato seleccionados para a amostra é proporcional ao número de elementos total existentes nesse estrato. Deste modo, a fração

de elementos na amostra mantém-se estável ao longo da população e a proporcionalidade do tamanho de cada estrato na população é mantida na amostra, garantindo que cada elemento da população tem a mesma probabilidade de pertencer a amostra. No entanto a variabilidade dentro de cada estrato não é considerada.

Várias vantagens advêm da aplicação deste plano de amostragem:

- Garante uma maior representatividade da amostra
- Dadas certas condições, a precisão das estimativas é maior relativamente a outros planos de amostragem (pois os erros padrão podem resultar do procedimento de estimação)
- Permite obter estimativas da variável de interesse para cada estrato, com uma dada precisão específica, assim como resolver problemas inerentes a cada um deles e que podem diferir de estrato para estrato
- Pode ser tão ou mais fácil recolher a informação neste tipo de amostragem como em a.a.s.. Se for o caso há pouco a perder tomando a estratificada pois os erros padrão raramente excedem os da a.a.s.

As principais dificuldades para a utilização desse tipo de amostragem residem nas complicações teóricas relacionadas com a análise dos dados e em que, muitas vezes, não podemos avaliar de antemão o desvio-padrão da variável nos diversos estratos.

2.3.1 Notação e Preliminares

Seja $\mathscr{P}$ a população de dimensão N (com N unidades) dividida em L estratos. Seja N_h o total de elementos da população no estrato h (h é o índice que identifica o estrato). Os estratos são disjuntos e a sua reunião coincide com toda a população, isto é, os estratos constituem uma partição de $\mathscr{P}$. Logo, tem-se que

$$N = \sum_{h=1}^{L} N_h. \tag{2.1}$$

Sendo n a dimensão da amostra, então n_h denota o número de elementos (unidades amostrais) do estrato h na amostra e

$$n = \sum_{h=1}^{L} n_h.$$

A relação entre a dimensão da população e a dimensão da amostra é representada pela fração amostral, $f = \frac{n}{N}$. Num determinado estrato h a fração é dada por $f_h = \frac{n_h}{N_h}$. Perante uma amostra aleatória estratificada através de afetação proporcional, as frações de amostragem são iguais em todos os estratos

$$f_h = f \Longrightarrow \frac{n_h}{N_h} = \frac{n}{N}. \tag{2.2}$$

Estimação de Proporções

As variáveis que serão consideradas posteriormente no estudo são qualitativas, sendo que a maioria delas nem são suscetíveis de ordenação. Por isso serão feitas inferências sobre proporções, onde a variável de interesse apenas tomará os valores 0 ou 1, sendo 1 caso o atributo em estudo seja observado no elemento amostral selecionado, e 0 caso o atributo nao esteja presente nesse elemento.

Define-se para este caso, a variável aleatória y_{hi}, que representa o valor da variável de interesse no elemento i do estrato h ($i = 1, ..., N_h$ e $h = 1, ..., L$):

$$y_{hi} = \begin{cases} 1, & \text{se o i-ésimo elemento do estrato } h \text{ tem o atributo} \\ 0, & \text{caso contrário} \end{cases} . \tag{2.3}$$

É com recurso a estas variáveis que se vai obter a expressão para a estimação da proporção P_h de elementos que possuem o atributo, no estrato h, para depois se estimar o mesmo em relação à população inteira, P.

Dentro de um determinado estrato h, como a amostragem é feita sem reposição, as variáveis y_{hi} correspondem a sucessões de experiências aleatórias dependentes e a probabilidade de cada elemento pertencer a amostra não é constante. Para além disso, as observações dividem-se em dois grupos, as que têm o atributo e as que não têm. Assim no estrato h de onde se recolhem n_h elementos, de um total de N_h elementos, existem K_h com o atributo e $N_h - K_h$ sem o atributo, e por isso o estimador para o total de elementos com o atributo tem distribuição hipergeométrica

$$\hat{\tau}_h = \sum\nolimits_{i=1}^{N_h} y_{hi} \sim H(N_h, n_h, K_h)$$

com valor médio e variância dados por

$$E(\hat{\tau}_h) = n_h(K_h/N_h) \quad \text{e} \quad Var(\hat{\tau}_h) = n_h(\frac{K_h}{N_h})(1 - \frac{K_h}{N_h})(\frac{N_h - n_h}{N_h - 1}).$$

O termo $(N_h - n_h)/(N_h - 1)$ corresponde a um fator de correção para populações finitas (neste caso a população corresponde ao estrato). A sua não aplicação aos estimadores terá como consequência uma sobre-estimação da estimativa da variância. Porém, verifica-se que este fator aproxima-se de 1 quando a dimensão da população é muito maior que a dimensão da amostra e por isso pode ser neglicenciado. Geralmente considera-se que a amostra é pequena em relação à população quando $\frac{n_h}{N_h} \leq 5\%$ ou até mesmo 10% [18]. Nesta situação, a probabilidade de um elemento ser selecionado varia muito pouco pois dificilmente um mesmo elemento será selecionado mais que uma vez. Consequentemente a amostragem com reposição vai produzir um resultado próximo ao da amostragem sem reposição, podendo assim ser usada a distribuição Binomial como aproximação para a distribuição Hipergeométrica.

Os parâmetros alteram-se então para

$$E(\hat{\tau}_h) = n_h P_h \quad \text{e} \quad Var(\hat{\tau}_h) = n_h P_h(1 - P_h) \quad \text{com} \quad P_h = \frac{K_h}{N_h},$$

pois

$$H(N_h, n_h, K_h) \approx B(n_h, P_h).$$

Nesta população, em que as observações são um conjunto de $0's$ e $1's$, verifica-se que a proporção é uma média destes valores. Assim, a proporção P_h corresponde ao valor médio do estrato h e a proporção amostral nesse estrato será intituitivamente um seu estimador. Note-se que nestes cálculos se pode considerar cada estrato como uma população em que a amostra foi selecionada por amostragem aleatória simples.

Estabelecem-se dois parâmetros ligados à população, tanto num estrato específico como em relação à população inteira:

Estrato h
Total de elementos da população com o atributo, no estrato h:

$$\tau_h = \sum_{i=1}^{N_h} y_{hi}.$$

Proporção de elementos da população com o atributo, no estrato h:

$$P_h = \frac{\tau_h}{N_h}.$$

População
Total de elementos da população com o atributo:

$$\tau = \sum_{h=1}^{L} \tau_h = \sum_{h=1}^{L} \sum_{i=1}^{N_h} y_{hi}.$$

Proporção de elementos da população com o atributo:

$$P = \frac{\tau}{N} = \frac{\sum_{h=1}^{L} \tau_h}{N} = \frac{1}{N} \sum_{h=1}^{L} N_h P_h.$$

Note-se que em amostragem estratificada, os parâmetros populacionais devem ser deduzidos ponderando a proporção do atributo em cada um dos estratos com o tamanho do respectivo estrato. Assim,

$$P = \frac{1}{N} \sum_{h=1}^{L} N_h P_h = \sum_{h=1}^{L} W_h P_h,$$

onde $W_h = N_H/N$ é o peso do estrato h (fração da população pertencente ao estrato h).

O estimador $\hat{P}$ é um estimador centrado, ou seja, $E(\hat{P}) = P$. A verificação assenta em cálculos simples. Por definição de P vem que

$$E(\hat{P}) = E\left(\frac{1}{N}\sum_{h=1}^{L} N_h \hat{P}_h\right) = \sum_{h=1}^{L} \frac{N_h}{N} E(\hat{P}_h),$$

onde

$$\hat{P}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h} \tag{2.4}$$

denota o estimador da proporção P_h no estrato h, proporção esta estimada através dos elementos da amostra. Note-se que os estratos são independentes e que em cada um deles, a amostra foi selecionada por amostragem aleatória simples, sem reposição.

Facilmente se verifica que $\hat{P}_h$, definido em (2.4), é um estimador centrado de P_h. Para o efeito, defina-se para cada estrato h a variável indicatriz I_{hi} tal que

$$I_{hi} = \begin{cases} 1, & \text{se } y_{hi} \in s_h \\ 0, & \text{se } y_{hi} \notin s_h \end{cases}, \ i = 1, \ldots, N_h, \tag{2.5}$$

onde y_{hi} são os definidos em (2.3) e s_h denota a amostra selecionada no h-ésimo estrato. Existe então uma v.a. de Bernoulli associada a cada elemento do estrato h (na população), com parâmetro π_{hi} igual à probabilidade de inclusão de y_{hi} na amostra s_h. Especificamente,

$$\pi_{hi} = P\{y_{hi} \in s_h\} = P\{I_{hi} = 1\} = \sum_{s_h \ni y_{hi}} P(s_h).$$

Foi referido na Secção 2.2 que, no caso do plano de amostragem aleatória simples sem reposição, $P(s_h) = \binom{N_h}{n_h}, \forall s_h \in \mathscr{S}_{n_h}$, onde $\mathscr{S}_{n_h}$ denota o espaço de todas as amostras possíveis de dimensão n. Tem-se então que

$$\pi_{hi} = P\{I_{hi} = 1\} = \sum_{s_h \ni y_{hi}} \frac{1}{\binom{N_h}{n_h}} = \frac{\binom{N_h-1}{n_h-1}}{\binom{N_h}{n_h}} = \frac{n_h}{N_h}.$$

Logo,

$$E(\hat{P}_h) = \frac{1}{n_h}\sum_{i=1}^{N_h} y_{hi}\, E(I_{hi}) = \frac{1}{n_h}\sum_{i=1}^{N_h} y_{hi}\pi_{hi} = \frac{1}{n_h}\,\frac{n_h}{N_h}\sum_{i=1}^{N_h} y_{hi} = \frac{K_h}{N_h} = P_h.$$

Portanto,

$$E(\hat{P}) = \sum_{h=1}^{L} \frac{N_h}{N} P_h = P.$$

Adicionalmente, a variância de $\hat{P}$ é dada por

$$Var(\hat{P}) = Var\left(\sum_{h=1}^{L} W_h \hat{P}_h\right) = \sum_{h=1}^{L} W_h^2 Var(\hat{P}_h).$$

Analogamente ao anterior, devido ao uso de variáveis aleatórias indicatrizes no contexto do plano de amostragem aleatória simples em cada estrato, obtém-se

$$Var(\hat{P}_h) = \left(\frac{N_h - n_h}{N_h - 1}\right) \frac{P_h(1 - P_h)}{n_h}$$

e assim

$$Var(\hat{P}) = \sum_{h=1}^{L} \frac{W_h^2}{n_h} \left(\frac{N_h - n_h}{N_h - 1}\right) P_h(1 - P_h). \tag{2.6}$$

Assumindo N_h suficientemente grande, de tal modo que $\frac{N_h}{N_h - 1} \to 1$, quando $N_h \to \infty$, o valor de $N_h - 1$ pode ser substituido por N_h, simplificando a expressão (2.6)

$$Var(\hat{P}) = \sum_{h=1}^{L} \frac{W_h^2}{n_h} \left(\frac{N_h - n_h}{N_h}\right) P_h(1 - P_h). \tag{2.7}$$

Como o valor de P_h não é conhecido, será estimado através das proporções amostrais do atributo em cada um dos estratos. Seja p a proporçao de elementos na amostra com o atributo, e p_h o mesmo mas referente ao estrato h. O estimador de P_h é p_h, isto é, $\hat{P}_h = p_h$. Logo,

$$\hat{P} = \frac{1}{N} \sum_{h=1}^{L} N_h \hat{P}_h = \frac{1}{N} \sum_{h=1}^{L} N_h p_h = \sum_{h=1}^{L} W_h p_h \tag{2.8}$$

com um estimador centrado da sua variância dado por

$$\widehat{Var}(\hat{P}) = \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h}\right) \frac{p_h(1 - p_h)}{n_h - 1}. \tag{2.9}$$

Estimação de Intervalos de confiança para Proporções

Na secção anterior obteve-se uma estimativa pontual para P, ou seja, um valor aproximado para o parâmetro que se desconhece. Este valor foi obtido a partir de uma dada amostra, de entre as várias amostras possíveis neste conjunto de dados. Qualquer uma outra amostra pode dar um valor diferente para a estimativa da proporção populacional, sem que se possa conhecer a magnitude do erro associado a cada uma destas estimativas, e assim não se tem como saber qual das estimativas será a melhor e que deve ser utilizada.

Perante esta limitação, uma alternativa consiste em procurar uma estimativa intervalar, e não pontual, para o parâmetro desconhecido. Pretende-se assim obter um intervalo de confiança (IC) para a estimativa da proporção, isto é, um intervalo com uma amplitude de

valores possíveis dentro dos quais pode estar o verdadeiro valor do parâmetro, com uma determinada confiança pré-definida. Um IC tem informação sobre a confiabilidade das estimativas, o que permite medir a precisão do estimador.

Já se sabe que o estimador $\hat{P}$ é centrado e tem variância estimada dada por (2.9). Se a dimensão de cada estrato fôr elevada ou o plano de amostragem tiver um grande número de estratos, o IC a $100 \times (1-\alpha)\%$ para a proporção referente à população inteira é dado por

$$\left]\hat{P} - z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{P})}\ ,\ \hat{P} + z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{P})}\right[. \tag{2.10}$$

Este intervalo apresenta uma confiança de $100 \times (1-\alpha)\%$, onde α representa uma probabilidade pequena associada ao erro cometido na obtenção do IC. O valor de $z_{1-\alpha/2}$ representa o quantil de probabilidade $1-\alpha/2$ da Normal(0,1) onde α depende da confiança que se pretende na construção do intervalo.

Consequentemente, a amplitude deste intervalo é definida por $2z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{P})}$.

2.4 Seleção da dimensão n da amostra

Depois de definida a população alvo a ser inquirida e escolhido o plano de amostragem mais adequado a esta população, o passo seguinte consiste na determinação da dimensão da amostra. Este ponto é bastante importante num estudo por amostragem, pois determina uma grande parte da fiabilidade dos dados. É necessário ter em conta aspectos como o nível de confiança e precisão desejados nos resultados, a variabilidade da população em estudo, os parâmetros que se pretendem estimar assim como os recursos de todo o tipo disponíveis. Pode-se concluir-se que a amostra deve ser o mais pequena possível, para evitar o desperdício de tempo e recursos, mas de tamanho suficiente para que se obtenham resultados válidos.

De um modo geral, na escolha da dimensão da amostra podem ser estabelecidas duas restrições: o gasto máximo que se pode fazer ou a precisão desejada. A expressão para o valor de n é conseguida tendo em conta um destes critérios. Neste estudo não houve nenhum tipo de limitações a nível de custos, apenas limitações de tempo, o que levou a ter-se escolhido definir a precisão desejada na obtenção de n.

Seja $\varepsilon > 0$ a precisão requerida na estimativa ou o erro de estimação tolerável. Este valor determina a diferença absoluta máxima admitida entre o valor real da proporção populacional e o valor estimado da proporção obtido da amostra, isto é, $\left|\hat{P} - P\right| \leq \varepsilon$. Considere-se também um nível de confiança α pequeno, que indica a probabilidade de o erro ε ser ultrapassado. Os valores que ε e α tomam devem ser previamente estabelecidos.

Pretende-se então determinar o valor mínimo n de tal modo que $\hat{P}$ seja um bom estimador de P com um precisão que não exceda ε e nível de confiança $1-\alpha$, isto é,

$$P\left(\left|\hat{P} - P\right| \leq \varepsilon\right) = \alpha.$$

Esta expressão é equivalente a

$$P\left(\right] \hat{P} - \varepsilon, \hat{P} + \varepsilon \left[\ni P\right) = \alpha. \tag{2.11}$$

o que traduz um intervalo a $100 \times (1-\alpha)\%$ de confiança, e assim $z\sqrt{Var(\hat{P})} = \varepsilon$. Substituindo a variância de $\hat{P}$ dada em (2.7) tem-se que

$$z\sqrt{\sum_{h=1}^{L} \frac{W_h^2}{n_h}\left(\frac{N_h - n_h}{N_h}\right) P_h(1-P_h)} = \varepsilon. \tag{2.12}$$

O parâmetro que se quer calcular, n, não está presente nesta equação e os valores n_h são desconhecidos, por isso é necessário encontrar alguma expressão que os relacione. Sabe-se que o número de elementos na amostra n_h de cada estrato h é uma fração da dimensão total da amostra n. Logo, tal como na população, existe um peso amostral w_h tal que $n_h = nw_h, i = 1, 2, \ldots, L$. A aplicação de afetação proporcional permite verificar, com recurso à fração amostral (2.2), que

$$w_h = \frac{n_h}{n} = \frac{N_h}{N} = W_h.$$

Introduzindo estes novos dados à equação (2.12) tem-se que

$$z\sqrt{\frac{1}{n}\sum_{h=1}^{L} W_h P_h(1-P_h) - \sum_{h=1}^{L} \frac{W_h^2}{N_h} P_h(1-P_h)} = \varepsilon. \tag{2.13}$$

Resolvendo em ordem a n e substituindo P_h por pelo seu estimador natural p_h, obtém-se a expressão pretendida,

$$n = \frac{\sum_{h=1}^{L} W_h p_h(1-p_h)}{(\varepsilon/z)^2 + \frac{1}{N}\sum_{h=1}^{L} W_h p_h(1-p_h)}. \tag{2.14}$$

Na prática o valor de P não é conhecido, pelo que pode ser substituido por uma estimativa. Como também não se dispõe de qualquer informação para obter esta estimativa, deve ser utilizado o valor da variância no pior caso, correspondendo à situação em que existe o máximo de heterogeneidade possível na população. A variância toma assim o valor de $1/2$ que corresponde à maior variabilidade populacional possível para a proporção de cada atributo. Esta atitude conduz a uma dimensão da amostra maior que o necessário mas que garante a precisão mínima imposta para a estimativa. De igual modo, para cada um dos estratos assume-se $\hat{P}_h = p_h = 1/2, h = 1, 2, \ldots, L$. Fazendo esta substituição em (2.14), obtém-se a seguinte expressão para o cálculo da dimensão n da amostra, para as condições deste estudo

$$n = \frac{N}{4N(\varepsilon/z)^2 + 1}. \tag{2.15}$$

2.5 Aplicação ao caso em estudo

Antes de aplicar o enquadramento teórico da secção 2.4 ao estudo que se pretende fazer, consideram-se divisões dos estratos diferentes para os particulares e empresas. Os particulares foram divididos por concelhos pois as características da população relativamente à formação podem diferir consoante a cidade, dado que pode haver locais com uma maior oferta de formações do que outros. Relativamente às empresas, considerou-se os setores de atividade económica existentes na Classificação Portuguesa de Atividades Económicas. Neste caso faz mais sentido dividir desta forma e não por concelhos, pois para uma dada empresa o seu comportamento em relação à formação que faz será semelhante a uma empresa dentro do mesmo tipo de actividade, e pode ser divergente da formação realizada numa empresa com uma actividade completamente diferente. Note-se que não foram considerados todos os setores pois em alguns deles, não faria sentido falar sobre formação devido ao seu tipo de actividade, e outros têm actividades dentro do mesmo género de setores escolhidos, podendo assim gerar conclusões muito semelhantes e tornar a análise dos resultados repetitiva.

Os particulares constituem a População 1 (notação:$\mathscr{P}_1$) e as empresas a População 2 (notação:$\mathscr{P}_2$). Para a $\mathscr{P}_1$ existem 19 estratos (concelhos) e um total de $N_{\mathscr{P}_1} = 627914$ residentes. Relativamente à $\mathscr{P}_2$, existem $N_{\mathscr{P}_2} = 22187$ empresas divididas por 7 estratos (setores).

Depois de definidos os estratos, o passo seguinte é o cálculo da dimensão da amostra. Considera-se um grau de confiança fixado em 95% ($z^2_{0.05/2} = 1.96$) e um erro amostral de 5% nos resultados, que são os valores geralmente usados. Com estes dados obtêm-se as dimensões da amostra $n_{\mathscr{P}_1}$ e $n_{\mathscr{P}_2}$, para os Partiulares e para as Empresas, respectivamente:

$$n_{\mathscr{P}_1} = \frac{N_{\mathscr{P}_2}}{4N_{\mathscr{P}_2}(\varepsilon/z)^2 + 1} = \frac{627914}{4 \times 627914 \times (0.05/1.96)^2 + 1} \approx 384,$$

$$n_{\mathscr{P}_2} = \frac{N_{\mathscr{P}_1}}{4N_{\mathscr{P}_1}(\varepsilon/z)^2 + 1} = \frac{22187}{4 \times 22187 \times (0.05/1.96)^2 + 1} \approx 378.$$

Já com os valores das dimensões das amostras, obtêm-se agora as frações amostrais respectivas para cada população:

$$f_{\mathscr{P}_1} = \frac{n_{\mathscr{P}_1}}{N_{\mathscr{P}_1}} = \frac{384}{627914} = 0.00061,$$

$$f_{\mathscr{P}_2} = \frac{n_{\mathscr{P}_2}}{N_{\mathscr{P}_2}} = \frac{378}{22187} = 0.0170.$$

Através de (2.2), calcula-se o número de elementos que devem ser selecionados em cada um dos estratos, n_h e que vão por fim constituir a amostra (uma para os particulares e outra para as empresas). Nas tabelas 2.1 e 2.2 encontram-se os valores obtidos.

Estrato h	Localidade	N_h	n_h
1	Águeda	24252	15
2	Albergaria-a-Velha	13819	9
3	Anadia	19999	12
4	Arouca	61170	37
5	Aveiro	126389	77
6	Castelo de Paiva	18491	11
7	Espinho	21126	13
8	Estarreja	42968	26
9	Ílhavo	22565	14
10	Mealhada	27290	17
11	Murtosa	61255	37
12	Oliveira de Azeméis	23977	15
13	Oliveira do Bairro	35602	22
14	Ovar	19408	12
15	Santa Maria da Feira	8215	5
16	São João da Madeira	20271	12
17	Sever do Vouga	49446	30
18	Vagos	10844	7
19	Vale de Cambra	20827	13

Tabela 2.1: Totais na amostra por estrato para particulares

Estrato h	Setor de Atividade Económica	N_h	n_h
C	Indústrias transformadoras	8080	137
G	Comércio por grosso e a retalho	8293	141
H	Transportes e armazenagem	1570	27
J	Actividades de informação e de comunicação	839	14
K	Actividades financeiras e de seguros	929	16
L	Actividades imobiliárias	1323	23
M	Actividades de consultoria, científicas, técnicas e similares	1153	20

Tabela 2.2: Totais na amostra por estrato para empresas

2.6 Estimação de uma Proporção

Considere-se a seguinte variável aleatória, correspondendo à definição apresentada em (2.3), com $h = 1, \ldots, L$ e $i = 1, \ldots, n_h$.

$$y_{hi} = \begin{cases} 1, & \text{se o i-ésimo elemento do estrato } h \text{ fez formação} \\ 0, & \text{se o i-ésimo elemento do estrato } h \text{ não fez formação,} \end{cases}$$

O objetivo é obter uma estimativa para a proporção de pessoas que fizeram formação, no caso dos Particulares, e uma estimativa da proporção de empresas que dão formação aos seus colaboradores, em relação às Empresas. Também se pretende determinar a variância estimada e um intervalo de confiança para estas estimativas da proporção. Estes 3 parâmetros são calculados com recurso a (2.8), (2.9) e (2.10), respetivamente. Note-se que o estimador da proporção em cada estrato, $\hat{P}_h$, é dado pela proporção amostral nesse mesmo estrato, p_h, sendo que $p_h = n_h^{-1} \sum_{i=1}^{n_h} y_{hi}$.

Os estratos nos particulares foram agrupados em 4 grupos, formando-se 4 novos estratos (motivo na Secção 3.3), $L = 4$, e a amostra tem $n_{\mathscr{P}_1} = 142$ elementos (Tabela 2.3).

Estrato h	$\sum_{i=1}^{n_h} y_{hi}$	n_h	N_h
Litoral Norte	11	19	92997
Interior Norte	18	30	149934
Litoral Centro	41	62	277840
Interior Centro	22	31	107143
Total	92	142	627914

Tabela 2.3: Tabela com os totais de respostas Sim, nos Particulares

Para as empresas definiram-se 7 estratos, ou seja, $L = 7$, tendo-se uma amostra de dimensão $n_{\mathscr{P}_2} = 230$ (Tabela 2.4).

Estrato h	$\sum_{i=1}^{n_h} y_{hi}$	n_h	N_h
C Indústrias transformadoras	83	83	8080
G Comércio por grosso e a retalho	75	86	8293
H Transportes e armazenagem	16	16	1570
J Actividades de informação e de comunicação	9	9	839
K Actividades financeiras e de seguros	10	10	929
L Actividades imobiliárias	9	14	1323
M Actividades de consultoria, científicas, técnicas e similares	12	12	1153
Total	214	230	22187

Tabela 2.4: Tabela com os totais de respostas Sim, nos Particulares

Note-se que os valores da dimensão das amostras são diferentes dos valores anteriormente calculados, devido ao número de observações que foram possíveis obter.

Pelos resultados que constam na Tabela 2.5 conclui-se que 64% da população frequentou alguma formação pelo menos uma vez. Relativamente às empresas esta percentagem aumenta para 93%, ou seja, quase a totalidade das empresas faz formação. O intervalo de confiança que possivelmente contém o verdadeiro valor da proporção nos Particulares tem uma amplitude maior que o dobro da amplitude do IC obtido para a proporção nas Empresas. Isto advém do facto de o erro padrão das observações nos Particulares ser maior, podendo ser considerado elevado, ao contrário do que acontece relativamente às Empresas. Assim, nas Empresas existe uma maior precisão da estimativa determinada.

	$\hat{P}$	$\widehat{Var}(\hat{P})$	Erro Padrão	IC a 95%	Amplitude IC
Particulares	0.643	0.00169	0.041	$]0.562, 0.723[$	0.080
Empresas	0.931	0.00024	0.016	$]0.900, 0.962[$	0.031

Tabela 2.5: Tabela das estimativas obtidas da Proporção, Variância e IC para Particulares e Empresas

Capítulo 3

Tabelas de Contingência

3.1 Tabelas de Contingência para 2 Factores

As tabelas de contingência são uma maneira de representar adequadamente informação cruzada sobre dados que são classificados por duas ou mais características. Geralmente, são aplicadas quando se está perante dados qualitativos, podendo também representar variáveis quantitativas se os seus valores forem discretizados ou agrupados em classes. Estas tabelas são uma ferramenta indispensável na análise estatística de variáveis categóricas pois permitem estudar a relação existente entre as suas variáveis, de forma relativamente rápida e precisa. As tabelas mais simples, que são aquelas que apresentam apenas duas variáveis, são tabelas bidimensionais.

Considerem-se dois atributos ou características representados pelas variáveis aleatórias X e Y, com r e s categorias, respetivamente. A tabela de contingência associada a uma amostra aleatória de X, cruzada com a caraterística Y, será uma tabela bidimensional com r linhas e s colunas, em que cada célula $(i, j), i = 1, ..., r, j = 1, ..., s$ contém a frequência n_{ij} das observaçõs amostrais classificadas simultaneamente com a caraterística i de X e j de Y:

	Y_1	Y_2	$\dots$	Y_s	Total
X_1	n_{11}	n_{12}	$\dots$	n_{1s}	$n_{1\cdot}$
X_2	n_{21}	n_{22}	$\dots$	n_{2s}	$n_{2\cdot}$
$\vdots$	$\dots$	$\dots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\dots$	n_{rs}	$n_{r\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\dots$	$n_{\cdot s}$	n

Tabela 3.1: Tabela de Contingência de 2 fatores, em termos de frequências absolutas

onde $n_{i\cdot}$ são os totais marginais da linha i e $n_{\cdot j}$ são os totais marginais da coluna j.

Podem ser associadas três tipos de probabilidades a uma tabela de contingência: con-

juntas, marginais e condicionais. Se cada observação for retirada da população através de um processo de amostragem aleatória simples e classificada em duas variáveis, tem-se a distribuição de probabilidade conjunta das variáveis. Esta distribuição é representada pelas probabilidades $\pi_{ij} = P(X = i, Y = j)$, que satisfazem $\pi_{ij} \geq 0$ e $\sum_{i,j} \pi_{ij} = 1$. Os totais das probabilidades conjuntas por linha, $\pi_{i\cdot} = \sum_j \pi_{ij}$, e por coluna, $\pi_{\cdot j} = \sum_i \pi_{ij}$, constituem a distribuição de probabilidade marginal de X e Y, respectivamente.

As correspondentes distribuições amostrais obtêm-se convertendo as contagens apresentadas na Tabela 3.1 em frequências relativas, dando o estimador das probabilidades conjuntas

$$\hat{\pi}_{ij} = p_{ij} = \frac{n_{ij}}{n}. \tag{3.1}$$

Estas proporções para cada uma das células fornecem a distribuição conjunta amostral e por isso estimam as probabilidades conjuntas π_{ij}. Do mesmo modo, calculam-se as distribuições marginais amostrais através de

$$\hat{\pi}_{i\cdot} = p_{i\cdot} = \sum_j p_{ij} = \frac{n_{i\cdot}}{n} \qquad \text{e} \qquad \hat{\pi}_{\cdot j} = p_{\cdot j} = \sum_i p_{ij} = \frac{n_{\cdot j}}{n}. \tag{3.2}$$

A Tabela 3.1 pode então ser reescrita em termos de proporções na forma da Tabela 3.2.

	Y_1	Y_2	$\ldots$	Y_s	Total
X_1	p_{11}	p_{12}	$\ldots$	p_{1s}	$p_{1\cdot}$
X_2	p_{21}	p_{22}	$\ldots$	p_{2s}	$p_{2\cdot}$
$\vdots$	$\ldots$	$\ldots$	$\ddots$	$\vdots$	$\vdots$
X_r	p_{r1}	p_{r2}	$\ldots$	p_{rs}	$p_{r\cdot}$
Total	$p_{\cdot 1}$	$p_{\cdot 2}$	$\ldots$	$p_{\cdot s}$	1

Tabela 3.2: Tabela de Contingência de 2 fatores, em termos de proporções

No caso de uma das variáveis ser uma variável resposta (ou dependente), sendo a outra uma variável explanatória, deve-se trabalhar em termos de probabilidade condicional. Neste caso de uma margem fixa (totais marginais à direita da Tabela 3.1 fixos), cada linha constitui uma amostra da distribuição multinomial, com probabilidades de categoria dados por $\pi_{j|i} = P(Y = j|X = i)$.

Estas probabilidades são estimadas através de

$$\hat{\pi}_{j|i} = p_{j|i} = \frac{n_{ij}}{n_{i\cdot}}. \tag{3.3}$$

ou seja, estuda-se a distribuição da variável resposta, seja Y essa variável, para cada uma das categorias da variável explanatória X.

3.1.1 Teste de Independência do Qui-Quadrado

O principal interesse em elaborar uma tabela de contingência consiste no posterior estudo que é possível fazer, relativamente à existência ou não de algum tipo de relação entre as suas variáveis. As variáveis podem ser independentes uma da outra ou, pelo contrário, pode existir associação (ou correlação) entre elas, com um determinado grau de associação que pode ser medido e posteriormente estimado. Se duas variáveis, X e Y, são independentes verifica-se a seguinte relação para todas as combinações possíveis de i e j

$$P(X=i, Y=j) \;=\; P(X=i)P(Y=j),$$

que é equivalente na amostra à igualdade (ver (3.1))

$$p_{ij} = p_{i\cdot}p_{\cdot j}.$$

Logo, sob a hipótese de independência, o valor esperado de observações com as características i de X e j de Y, de entre as n observações, é dado por

$$e_{ij} = np_{ij} = np_{i\cdot}p_{\cdot j} = n\frac{n_{i\cdot}}{n}\frac{n_{\cdot j}}{n} = \frac{n_{i\cdot}n_{\cdot j}}{n}. \tag{3.4}$$

Calcula-se o valor esperado para todas as células da tabela, para depois se inferir através de um teste de hipóteses se as variáveis são ou não independentes. Testa-se assim as seguintes hipóteses alternativas:

$$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}, \forall(i,j) \quad vs. \quad H_1 : \pi_{ij} \neq \pi_{i\cdot}\pi_{\cdot j}, \text{para algum } (i,j). \tag{3.5}$$

Isto significa que a independência entre as variáveis existe se e só se a frequência observada de cada uma das células da tabela fôr igual ao produto da frequência total da linha correspondente pela frequência total da coluna correspondente, dividido pelo total de observações. Esta hipótese considera que as duas variáveis são independentes uma da outra se as suas probabilidades conjuntas amostrais, divididas pelo total de observações, forem iguais às frequências esperadas em casod de independência.

Para tratar as variáveis apresentadas, que são nominais, será utilizado o Teste do Qui-Quadrado com a estatística de teste definida por

$$X^2 := \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}.$$

Esta espressão é também designada por estatística de Pearson e pela forma como está definida, a estatística de teste mede o afastamento dos dados em relação à hipótese de independência, verificando a concordância entre as frequências observadas e as frequências esperadas.

Assumindo a hipótese nula como verdadeira, pode-se considerar que esta estatística de teste tem distribuição assintótica Qui-Quadrado com $(r-1)(s-1)$ graus de liberdade, i.é.,

$$X^2 \sim \chi^2(r-1)(s-1).$$

A hipótese nula de independência deverá ser rejeitada quando o valor da estatística de teste é maior que um determinado valor crítico, obtido a partir da função quartil da distribuição Qui-Quadrado, para um dado nível de significância α escolhido. Ou seja,

$$X^2 \geq \chi^2_{1-\alpha}(r-1)(s-1).$$

Esta expressão significa que para valores elevados da estatística de teste, que ultrapassam o valor crítico da distribuição χ^2, a hipótese de independência das variáveis deve ser rejeitada. Isto acontece quando as frequências observadas são substancialmente diferentes das frequências esperadas sob independência. Tem-se assim a região crítica do teste, associado ao nível de significância α,

$$\mathrm{RC}_\alpha = \{X^2 : X^2 > \chi^2_{1-\alpha}(r-1)(s-1)\}.$$

O valor p (*p-value*) do teste é

$$p - value = P(X^2 \geq X^2_{obs}|H_0).$$

rejeitando-se a hipótese nula se $p - value \leq \alpha$. Fixa-se desde já o nível de significância dos testes a serem realizados em $\alpha = 0.05$.

Havendo evidências estatísticas em como se deve rejeitar a hipótese nula de independência, pode interessar medir o grau de dependência entre as variáveis, através de uma medida de associação apropriada, assim como saber qual a fonte de dependência que deu origem à rejeição da hipótese nula.

O uso da distribuição do Qui-Quadrado como uma aproximação para a distribuição da estatística de teste X^2 é válido sob a hipótese de que as frequências esperadas não são muito pequenas e as amostras suficientemente grandes. O teste de independência do Qui-Quadrado deve assim ser aplicado apenas se se verificam determinados pressupostos, aqui descritos em termos da "regra do polegar" [14]:

- Se $n \leq 20$, nenhuma célula da tabela deve ter frequência esperada inferior a 5 unidades.
- Se $n > 20$, não deverá haver mais do que 20% das células com frequência esperada inferior a 5 nem alguma célula com frequência esperada inferior a 1.

Caso os valores das frequências esperadas sejam pequenos, os valores possíveis para a estatística X^2 serão bastante discretos o que torna a aproximação pela distribuição contínua do Qui-Quadrado inválida [9].

Uma alternativa a este teste de independência é o teste exato de Fisher, que calcula o valor exato do p-value. Outra solução para este problema consiste em agrupar algumas categorias das variáveis. No entanto, esta solução pode originar uma perda de informação dos dados, podendo consequentemente provocar efeitos na utilidade e viabilidade do estudo, assim como nas inferências que seriam obtidas. Com esta técnica também se perde a aleatoriedade da amostra, com consequências desconhecidas.

Independência em dados ordinais

Os testes acima referidos não têm em conta situações em que os dados estão ordenados. Nestes casos é comum haver uma tendência linear na associação, positiva ou negativa. Assim, as estatísticas de teste aplicadas devem tratar as variáveis ordinais como quantitativas em vez de qualitativas, providenciando um maior poder no teste. Geralmente, esta análise é efetuada atribuindo *scores* às várias categorias ordenadas e medindo o grau de tendência linear.

Seja $u_1 \leq u_2 \leq \ldots \leq u_r$ e $v_1 \leq v_2 \leq \ldots \leq v_s$ os *scores* escolhidos para atribuir às linhas e colunas, respectivamente, e com a mesma ordem das correspondentes categorias ordenadas. Tomando $\bar{u} = \sum_i u_i p_{i+}$ a média amostral dos *scores* das linhas e $\bar{v} = \sum_j v_j p_{+j}$ a média amostral dos *scores* das colunas, o coeficiente de correlação de Pearson r entre as variáveis X e Y de uma tabela de contigência como a apresentada em 3.1 é dado por

$$r = \frac{\sum_{i,j}(u_i - \bar{u})(v_j - \bar{v})p_{ij}}{\sqrt{\sum_i (u_i - \bar{u})^2 p_{i+} \ \sum_j (v_j - \bar{v})^2 p_{+j}}}$$

O valor de r varia entre -1 e 1, onde a independência acontece quando $r \neq 0$. Este valor mede a força e a direção da associação linear. Quanto maior o seu valor absoluto, maior distância existe da tendência linear.

Deste modo, pretende-se testar a hipótese nula de independência $H_0 : r = 0$, isto é, de que não existe associação linear entre as linhas e colunas da tabela. A estatística de teste para este teste é

$$M = (n-1)r^2 \tag{3.6}$$

cuja distribuição assintótica, quando n é suficientemente grande, é a distribuição Qui-Quadrado com 1 grau de liberdade. [10] [21]

3.1.2 Teste de Homegeneidade do Qui-Quadrado

Quando a população é dividida em r subpopulações ou estratos, $S_1, S_2, \ldots, S_r$, e se pretende saber o seu comportamento nas categorias de uma dada variável, aplica-se um teste de homogeneidade em vez de um teste de independência. Deste modo, as tabelas de contingência têm uma das margens fixa (sejam os totais por linha ou os totais por coluna), pois o número de observações a retirar aleatoriamente de cada subpopulação é previamente fixado. Em vez de probabilidades conjuntas, utilizam-se aqui probabilidades condicionais.

Pretende-se assim saber se o modo como as subpopulações se distribuem pelas várias categorias da variável resposta é homogénea e a hipótese de teste para verificar se existe homogeneidade na amostra expressa-se como:

$$H_0 : \pi_{j|1} = \pi_{j|2} = \ldots = \pi_{j|r}, j = 1, ..., c. \tag{3.7}$$

Caso haja homogeneidade, é de esperar que a proporção de ocorrências numa dada categoria da variável resposta seja a mesma para todas as subpopulações.

Neste contexto a região de rejeição e a regra de decisão são as mesmas do teste de independência, apesar de as conclusões que se retiram num teste de homogeneidade serem diferentes.

3.1.3 Análise de resíduos

Em caso de rejeição da hipótese nula, pode interessar descobrir quais as classificações cruzadas da tabela de contingência que provocaram esta rejeição. Para isso comparam-se os desvios entre as frequências observadas e as frequências esperadas, célula a célula, de modo a medir o afastamento da hipótese de independência em cada uma delas, isto é, o quanto contribuiram para a estatística X^2. Este afastamento é analisado através dos chamados resíduos

$$r_{ij} = \frac{o_{ij} - e_{ij}}{\sqrt{e_{ij}}},$$

cuja estimativa da variância é dada por

$$v_{ij} = (1 - \frac{n_{i\cdot}}{n})(1 - \frac{n}{n}).$$

Normalmente utilizam-se os resíduos reduzidos, designados por d_{ij}, pois estes têm distribuição assintótica normal padronizada,

$$d_{ij} = \frac{r_{ij}}{\sqrt{v_{ij}}} \sim N(0,1).$$

Esta aproximação foi demonstrada por Haberman (1973) [8]. Comparando os valores dos resíduos com o quantil de probabilidade $1 - \alpha/2$, percebe-se quais os resíduos significativos, para o nível de significância α. Se um resíduo d_{ij} de uma dada célula (i, j) for significativo, isto é, $|r_{ij}/\sqrt{v_{ij}}| > z_{1-\alpha/2}$, então essa célula contribuiu significativamente para a estatística X^2. Logo a classificação correspondente é uma das responsáveis por se assumir a dependência das variáveis.

3.1.4 Medidas de associação

O valor da estatística X^2 é determinante para a rejeição ou não da hipótese de independência, mas não dá informação alguma relativamente ao grau de dependência das variáveis, caso se rejeite H_0. Para isso, existem medidas de associação que permitem quantificar este grau de dependência.

De entre as várias medidas de associação existentes, nem todas podem ser aplicadas a tabelas onde as variáveis tem mais de 2 categorias, que é precisamente o que acontece neste trabalho. Para estas tabelas existem pelo menos duas medidas mais usadas: o coeficiente de contingência de Pearson e o coeficiente de Cramér. No entanto, estas medidas não têm uma interpretação significante no sentido em que não dão informação sobre o modo como as variáveis estão associadas, nem se a relação é forte, interessante ou relevante, não acrescentando assim informação útil às conclusões. Outra desvantagem é que não podem

ser interpretadas probabilisticamente em termos dos dados da tabela. Por exemplo, caso a formação realizada esteja eventualmente associada ao género não parece ser relevante dizer se esta associação é alta (forte) ou não, apenas interessa saber que a dependência existe.

Outro modo de avaliar o grau de associação entre duas variáveis é através da Razão das Possibilidades (*Odds ratio*). Esta medida tem características um pouco diferentes das citadas anteriormente e será a única aplicada neste estudo para medir a força da associação entre as variáveis. Normalmente é usada em tabelas de contingência 2×2 mas pode ser generalizada para subconjuntos 2×2 de tabelas de maiores dimensões. A sua principal vantagem é que se trata de uma medida que não é afetada pela dimensão da amostra nem pelos totais marginais. Depende simplesmente da associação entre as variáveis.

Define-se primeiro o conceito de Possibilidade (*Odds*), que se define como a razão entre a probabilidade de um acontecimento e a probabilidade do seu complementar. Sendo $\pi = P(A)$ a probabilidade de um acontecimento A, então a possibilidade é dada por

$$Odds = \text{Possibilidade} \equiv \frac{P(A)}{P(A^C)} = \frac{P(A)}{1 - P(A)} = \frac{\pi}{1 - \pi}.$$

Por exemplo, se $odds = 3/2$ então para cada 3 sucessos são esperados 2 insucessos. Observa-se que $odds$=1 corresponde a $\pi = 0.5$, ou seja, o sucesso e o não sucesso são igualmente prováveis. Quando o sucesso é mais provável que o não sucesso, então $\pi > 0.5$ e $odds > 1$. Se o não sucesso é mais provável que o sucesso, $\pi < 0.5$ e $odds < 1$.

Relacionada com Possibilidade, tem-se a Razão das Possibilidades. Para uma variável resposta binária observada em dois grupos, numa tabela de contingência 2×2 de margens livres, sendo π_1 a probabilidade de sucesso do grupo 1 e π_2 a probabilidade de sucesso do grupo 2, esta razão é dado por

$$Odds\ ratio = \text{Razão das Possibilidades} \equiv \theta = \frac{\pi_1/(1 - \pi_1)}{\pi_2/(1 - \pi_2)}.$$

A razão das possibilidades pode ser utilizada em tabelas com uma margem fixa, tal como no caso de tabelas 2×2 de margens livres. Recorde-se que no caso de uma margem fixa, as frequências aí apresentadas correspondem a estimativas das probabilidades condicionais $\pi_{j|i}$. Assim sendo, basta verificar que a razão das possibilidade θ admite uma representação em termos destas probabilidades:

$$\theta = \frac{\pi_{11}\pi_{22}}{\pi_{21}\pi_{12}} = \frac{\frac{\pi_{1|1}}{\pi_{1\cdot}}\frac{\pi_{2|1}}{\pi_{2\cdot}}}{\frac{\pi_{1|2}}{\pi_{2\cdot}}\frac{\pi_{2|1}}{\pi_{1\cdot}}} = \frac{\pi_{1|1}\,\pi_{2|2}}{\pi_{1|2}\,\pi_{2|1}}.$$

Note-se que o valor $\theta = 1$ significa que ambos os grupos têm a mesma probabilidade de sucesso (e portanto a mesma probabilidade de insucesso), o que implica independência das variáveis que presidem à tabela. Caso $\theta > 1$, a categoria com índice 1 tem probabilidade de sucesso maior do que a categoria 2, i.e., $\pi_1 > \pi_2$. Caso $\theta < 1$, então $\pi_2 > \pi_1$. Assim sendo,

a independência entre as duas variáveis de interesse poderá ser detetada como significativa a partir de estimativas $\hat{\theta}$ suficientemente próximas de 1.

Adotando para estimador da razão das possibilidades o decorrente de substituir as probabilidades π pelas correspondentes probabilidades empíricas p (cf. (3.3) e (3.1)) disponíveis nas tabelas de contigência, tem-se:

$$\hat{\theta} := \frac{p_{11}p_{22}}{p_{21}p_{12}} = \frac{p_{1|1}\, p_{2|2}}{p_{1|2}\, p_{2|1}}, \tag{3.8}$$

sendo que a segunda igualdade acima se reporta ao caso de tabelas com uma margem fixa. A razão de possibilidades tem a mesma expressão tanto num contexto de amostragem com amostras binomiais independentes como em qualquer outro esquema de amostragem. Tal acontece porque esta medida não distingue entre variável resposta e variável explanatória. De agora em diante, será dado especial ênfase a este caso porque é o mais natural para a aplicação prática no âmbito do presente trabalho, com a amostragem estratificada.

Como θ e $\hat{\theta}$ são sempre não negativos, pois são respetivamente funcionais de proporções e de frequências empíricas, a distribuição teórica de $\hat{\theta}/\theta$ assume valores no intervalo $[0, \infty[$ sendo bastante assimétrica para vários valores de θ. Quando a dimensão da amostra n é suficientemente elevada, tem-se que a variável aleatória $\hat{\theta}/\theta - 1$, convenientemente normalizada, é assintoticamente normal mas existem comprovadas vantagens na utilização da transformação do tipo logaritmo para alcançar maior simetria. Por isso, a inferência sobre a razão das possibilidades $\theta \geq 0$ é geralmente realizada mediante a consideração da estatística $\log(\hat{\theta}/\theta)$. O método delta de Cramér permite então concluir que $\log\hat{\theta} - \log\theta - \log 1 = \log\hat{\theta} - \log\theta$, mediante normalização adequada, ainda é assintoticamente normal. Note-se que com esta transformação, a independência entre as variáves é assumida quando $\log\theta = \log 1 = 0$.

Deste modo, a sua distribuição pode ser aproximada pela distribuição normal, com média $\log\theta$ e desvio-padrão

$$\hat{\sigma}_{\log\hat{\theta}} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{22}} + \frac{1}{n_{12}} + \frac{1}{n_{21}}},$$

logo para n suficientemente grande, tem-se sob a hipótese de não associação (independência) das variáveis,

$$\frac{\log\hat{\theta}}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{22}} + \frac{1}{n_{12}} + \frac{1}{n_{21}}}} \sim N(0,1), \tag{3.9}$$

ou equivalentemente,

$$\frac{(\log\hat{\theta})^2}{n_{11}^{-1} + n_{22}^{-1} + n_{12}^{-1} + n_{21}^{-1}} \sim \chi^2(1). \tag{3.10}$$

Portanto a estatística em (3.10) poderá ser diretamente utilizada para testar $H_0 : \theta = 1$ *versus* $H_1 : \theta \neq 1$, sendo que a hipótese nula de independência deverá ser rejeitada para valores grandes desta estatística de teste. Exemplos de aplicação desta estatística são apresentados mais adiante em 3.4.1 e 3.4.3.

Note-se que em tabelas que contenham variáveis com mais de duas categorias, a razão de possibilidades também pode ser calculada, combinando quaisquer duas linhas com quaisquer duas colunas. Numa tabela com I linhas e J colunas, haveriam $\binom{I}{2}\binom{J}{2}$ razões de possibilidades possíveis, havendo contudo informação redundante. Utiliza-se apenas um subconjunto mínimo de $(I-1)(J-1)$ razões locais, cuja expressão geral para cada θ é

$$\theta_{ij} = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i,j+1}\pi_{i+1,j}}, \quad i = 1, ..., I-1, j = 1, ..., J-1. \tag{3.11}$$

Uma exposição mais detalhada sobre este tópico pode ser consultada em [1].

3.2 Tabelas de Contingência para 3 Fatores

Em tabelas tridimensionais, não existe apenas independência ou associação dos factores. Estas tabelas descrevem a relação entre 3 fatores, considerando tabelas de contingência com 2 fatores, chamadas de tabelas parciais, para cada categoria da terceira variável. Neste contexto, podem ser definidos vários tipos de independência possíveis de identificar entre os três factores. Considere-se as variáveis X, Y e Z, com r, c e l categorias, respectivamente.

1. **Independência Mútua [X][Y][Z]**
   Neste tipo de independência, não há qualquer associação entre as variáveis. Esta hipótese de independência completa formula-se da seguinte forma:

   $$H_0^{(1)} : \pi_{ijk} = \pi_{i..}\pi_{.j.}\pi_{..k}, \quad \forall i, j, k.$$

   Se se rejeitar H_0 conclui-se que existe associação significativa. Caso contrário, deve-se testar outro tipo de independência para obter alguma conclusão.

2. **Independência Conjunta [XY][Z]**
   A hipótese considerada é a de que o par de variáveis formado por X e Y é conjuntamente independente de Z, ou seja, tem-se um factor independente dos outros dois. Formalmente

   $$H_0^{(2)} : \pi_{ijk} = \pi_{ij.}\pi_{..k}, \quad \forall i, j, k.$$

   Caso se decida não rejeitar H_0, para além de se verificar a independência conjunta referida, esta vai implicar também a independência marginal tanto entre X e Z, como entre Y e Z. Sobre a relação existente entre X e Y nada se pode concluir sem um outro teste adequado.

3. **Independência Condicional [XZ][YZ]**
   Considera-se agora a independência das variáveis X e Y condicional à variável Z, ou de outra forma, as variáveis X e Y são independentes dada cada categoria da variável Z. Neste tipo de independência,

   $$H_0^{(3)} : \pi_{ijk} = \frac{\pi_{i.k}\pi_{.jk}}{\pi_{..k}}, \quad \forall i, j, k.$$

Na sequência da não rejeição de H_0, se houver alguma associação entre X e Y, esta desaparece quando se controla por uma terceira variável Z. Caso se rejeite H_0 interessa descobrir qual ou quais as categorias de Z que provocam o afastamento da independência. Adicionalmente à aceitação da hipótese, se se souber que X é independente de Z e Y é independente de Z então é imediato que existe independência mútua entre as três variáveis. Uma outra hipótese equivalente a H_0, neste tipo de independência é

$$H_0^{(3')} : \theta_{ij(1)} = \ldots = \theta_{ij(k)} = 1 \quad \forall i, j, k.$$

4. **Associação Homogénea [XY][XZ][YZ]**
   Este tipo de associação permite que todos os três pares de variáveis tenham associação condicional. A hipótese é estabelecida em termos da razão de possibilidades.

$$H_0^{(4)} : \theta_{ij(1)} = \ldots = \theta_{ij(k)}, \quad i = 1, ..., r-1, j = 1, ..., s-1, k = 1, ..., l.$$

   onde $\theta_{ij(k)} = \frac{\pi_{ij(k)}\pi_{i+1,j+1(k)}}{\pi_{i,j+1(k)}\pi_{i+1,j(k)}}$.
   A razão de possibilidades condicional entre cada par de variáveis é igual em cada nível da terceira variável, o que significa que a associação entre duas variáveis não depende das categorias da restante variável.

Para os modelos 2, 3 e 4, existem modelos análogos que se obtêm trocando as variáveis de posição. Os estimadores de máxima verosimilhança das probabilidades são as frequências relativas (proporções) correspondentes, usando as distribuições marginais, isto é, $\hat{\pi}_{ijk} = p_{ijk}$.

3.3 Caracterização geral da população: Particulares e Empresas

Apresenta-se inicialmente uma caracterização geral da parte da população que foi inquirida, ou seja, da amostra. Na Figura 3.1 a amostra está dividida por localidade, género e faixa etária. Observa-se que existem localidades onde não foram inquiridos homens e outras nas quais não se tem pessoas de uma dada faixa etária. Também se verifica que o número de elementos que constam na amostra, por estrato (localidade), é inferior ao valor obtido nos cálculos efectuados na Secção 2.5, para o respectivo estrato. Dada a natureza do presente estudo, a informação recolhida para as diversas categorias nos vários estratos é intrinsecamente esparsa. Isto significa que uma recolha exaustiva de dados é naturalmente impraticável. Portanto, de modo a tornar a distribuição dos dados mais homogénea, agrupou-se as localidades por regiões tendo em conta a sua localização geográfica.

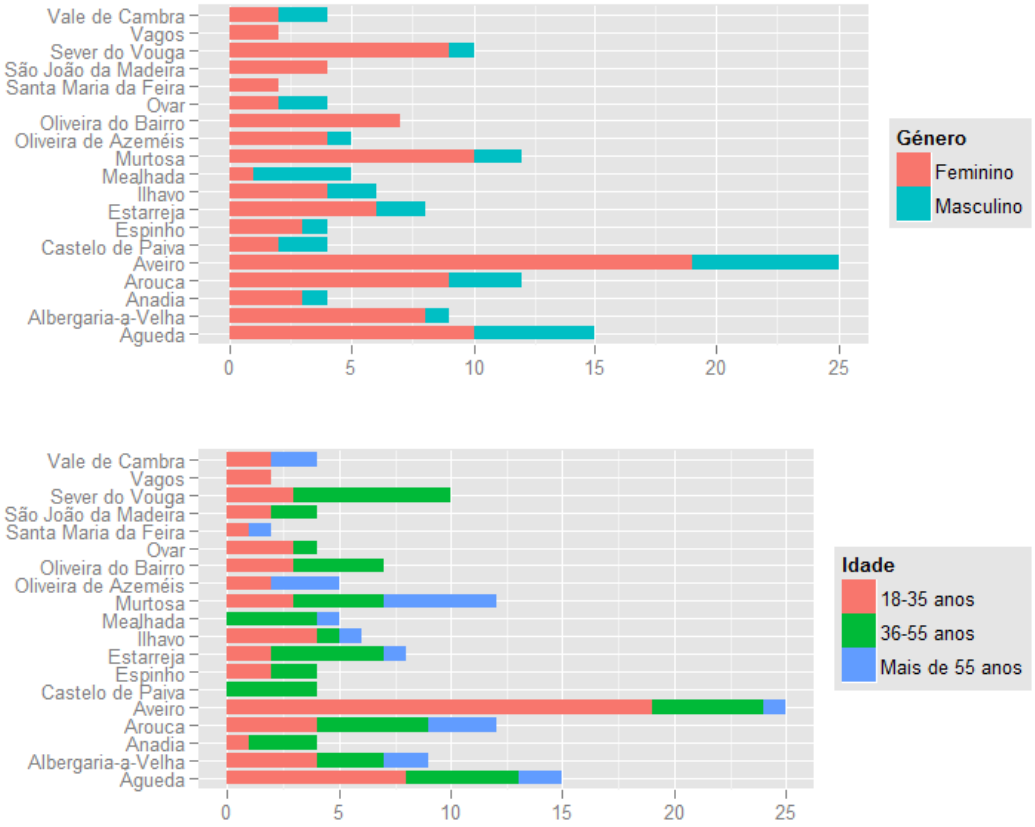

Figura 3.1: Amostra de Particulares nas Localidades, por Género e Idade

Assim, a população dos particulares será estudada através de quatro regiões:

- **Região do Litoral Norte:** Espinho, Santa Maria da Feira, São João da Madeira, Ovar, Oliveira de Azeméis
- **Região do Interior Norte:** Arouca, Castelo de Paiva, Sever do Vouga, Vale de Cambra
- **Região do Litoral Centro:** Albergaria-a-Velha, Aveiro, Estarreja, Ílhavo, Murtosa, Vagos
- **Região do Interior Centro:** Águeda, Anadia, Mealhada, Oliveira do Bairro

Os elementos amostrais por região estão caracterizadas na Tabela 3.3, por género e idade. Observa-se que a maioria da amostra é do sexo feminino e está na faixa etária dos $18 - 35$ anos. A faixa etária em menor representação na amostra é a dos maiores de 55 anos.

		18-35 anos	36-55 anos	+55 anos	Total Região
Litoral Norte	Feminino	8	4	3	19
	Masculino	2	1	1	
Interior Norte	Feminino	7	12	3	30
	Masculino	2	4	2	
Litoral Centro	Feminino	26	15	8	62
	Masculino	8	3	2	
Interior Centro	Feminino	10	11	0	31
	Masculino	2	5	3	

Tabela 3.3: Frequências amostrais observadas dos Particulares, nas Regiões, por Género e Idade

Relativamente às empresas, a Figura 3.2 mostra como as empresas que acederam responder ao inquérito se distribuem pelas localidades do Distrito de Aveiro. Duas localidades destacam-se neste gráfico relativamente ao total de empresas inquiridas, Santa Maria da Feira e Aveiro. Em Aveiro talvez por ser a capital de distrito, onde se centralizam bastantes empresas e serviços, nas mais diversas áreas. Em relação a Santa Maria da Feira, é uma zona com grande actividade industrial, dai ter também um elevado número de empresas. Assim, há uma maior probabilidade de as empresas escolhidas aleatoriamente dentro de cada estrato (setor) serem destas duas localidades. Já Castelo de Paiva foi a única localidade em que nao houve registo de empresas que tenham respondido ao inquérito.

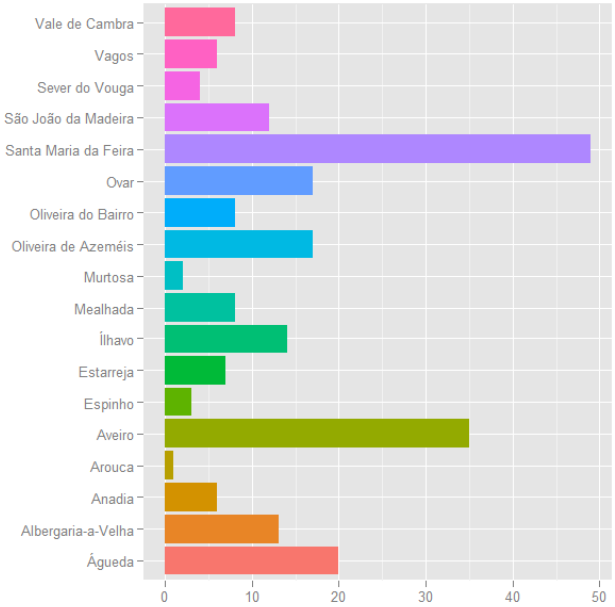

Figura 3.2: Amostra de Empresas por Localidade

3.4 Análise de relações entre questões: Particulares

Os dados recolhidos foram organizados em tabelas de contingência com 2 e 3 fatores, às quais foram aplicados testes de independência, com o objetivo de averiguar o tipo de independência existente entre as variáveis da tabela. Nas secções seguintes, os testes de indpendência do Qui-Quadrado efetuados pretendem testar a hipótese de independência 3.5 entre as variáveis em questão em cada secção. Numa primeira fase foram aplicados a pares de variáveis, nos quais uma delas é sempre a variável Formação, com exceção de um par. As associações testadas foram definidas de acordo com uma linha de pensamento em que foi ponderado se, para uma determinada variável, faria algum sentido relacioná-la com a formação. Sempre que em algum desses testes uma das variáveis em causa seja a Região, o teste realizado é o de homogeneidade, que corresponde ao teste de hipóteses 3.7. Numa segunda fase foram examinadas tabelas com 3 variáveis, pois a cada par foi adicionada a variável Região.

Antes de efetuar qualquer teste de hipóteses, foram examinadas as frequências esperadas referentes às observações das tabelas às quais se vai aplicar o teste. Depois de calculadas com recurso ao comando *chisq.test()$expected* do programa R, averiguou-se se satisfazem as regras da Subsecção 3.1.1. Caso estas regras se verifiquem, é possível usar o teste de independência do Qui-Quadrado, através da função *chisq.test()*. Caso contrário, deve antes ser aplicado o teste exato de Fisher, através de *fisher.test()*.

3.4.1 Género - Idade - Formação

Inicia-se o estudo dos particulares com a análise das variáveis Género e Idade de modo a verificar se estas duas variáveis têm alguma relação/influência na formação.

Género	Formação	Idade			Total Formação	Total Género
		18-35 anos	36-55 anos	+55 anos		
Feminino	Sim	40	29	5	74	107
	Não	11	13	9	33	
Masculino	Sim	8	8	2	18	35
	Não	6	5	6	17	
Total Idade		65	55	22	142	

Tabela 3.4: Tabela de contingência $2 \times 3 \times 2$ das variáveis Formação, Género e Idade

Para o par de variáveis Género e Formação pode ser usado o teste de independência do Qui-Quadrado, já que todas as frequências esperadas da tabela correspondente são maiores que 5. O valor observado da estatística de Pearson é

$$X^2_{obs} = 2.8986.$$

O número de graus de liberdade (g.l.) da tabela associada a estas duas variáveis (sub-tabela da Tabela 3.4 apenas com as variáveis), cada uma com 2 categorias, é $(r-1)(c-1) = (2-1)(2-1) = 1$. O valor que, numa distribuição χ^2 com 1 g.l., deixa à direita uma região de probabilidade $\alpha = 0.05$, i.e., o quartil de probabilidade $1-\alpha = 0.95$ da distribuição $\chi^2_{0.95}(1)$ é:

$$\chi^2_{0.95}(1) \cong 3.8415.$$

Comparando estes valores, $X^2_{obs} < \chi^2_{0.95}(1)$, logo o valor observado não pertence à região crítica do teste ao nível de significância $\alpha = 0.05$. Assim, ao nível 5%, não há evidências para rejeitar a hipótese de independência, admitindo-se que não existe associação entre o género de uma pessoa e a formação realizada. A mesma conclusão poderia ser obtida verificando que o *p-value* do teste é $0.089 > 0.05 = \alpha$.

Em relação à variável Idade, têm-se 3 categorias (Tabela 3.3 que motiva a classificação na Tabela 3.4). Porém, esta variável aqui dividida por faixas etárias, é naturalmente uma variável ordenada. Logo deve ser usado um teste adequado como o descrito na Subsecção 3.1.1, de deteção de tendência. Os seguintes *scores* foram atribuidos às variáveis em causa:

Idade	18-35 anos: $u_1 = 1$	Formação	Sim: $v_1 = 1$
Covariável	36-55 anos: $u_2 = 2$	Var. dependente	Não: $v_2 = 0$
	+55 anos: $u_3 = 2$		

Tanto o valor de r como o valor da estatística M (Eq. (3.6)) foram calculados com recurso à função *pears.cor()*, cujos resultados foram $r = -0.27$ e $M = 10.160$. O valor do coeficiente de correlação sugere que a associação linear é negativa, isto é, à medida que se avança nas faixas etárias, o nível de formação vai diminuindo. A estatística de teste M com 1 grau de liberdade tem $p - value = 0.001$, que também sugere uma forte evidência de uma correlação não nula.

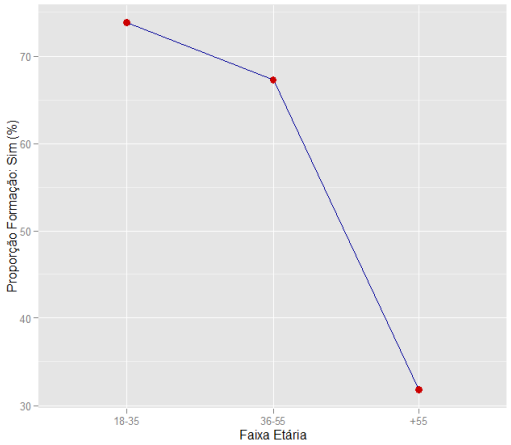

Figura 3.3: Proporção de Formação realizado em cada faixa etária da variável Idade

No gráfico da Figura 3.3 observa-se que efetivamente a proporção de formação realizada diminui, ao longo das 3 faixas etárias ordenadas por ordem crescente. Esta observação faz sentido pois é normal que uma pessoa com idade entre os 18 e os 35 se encontre no final do ensino obrigatório e início de vida profissional, e portanto é lógico que frequente formação. A faixa etária seguinte ainda engloba indivíduos com idade de trabalhar logo também é normal a frequência de formação, observando-se no entanto uma menor incidência. Em idade superior a 55 anos, é natural que a disponibilidade para obter novos conhecimentos, ou melhorar os que já se possui, seja menor e daí o visível decréscimo de proporção.

Neste caso, devem ser averiguadas quais as classificações responsáveis pela não aceitação da independência entre as variáveis. Para isso, usam-se os resíduos, que ajudam na interpretação das conclusões. Na seguinte tabela encontram-se os resíduos reduzidos associados à Tabela 3.4, colapsando-a sobre a variável Género.

	Sim	Não
18-35 anos	2.076	-2.076
36-55 anos	0.493	-0.493
+55 anos	-3.522	3.522

Tabela 3.5: Resíduos padonizados entre a Idade e a Formação

A tabela tem grandes resíduos positivos (> 2) para as pessoas entre os 18 e 35 anos que fizeram formação e para pessoas com mais de 55 anos que não fizeram. Significa que existem menos pessoas nestas categorias cruzadas que o que seria de esperar sob a hipótese de independência. Os maiores resíduos negativos correspondem exatamente ao contrário dos resíduos positivos, isto é, pessoas entre os 18 e 35 anos que não fizeram formação e para pessoas com mais de 55 anos que fizeram. Seria de esperar em caso de independência que houvesse mais observações nestas células.

Resumindo, se as variáveis fossem independentes, esperar-se-ia para a faixa etária $18-55$ anos que houvesse menos pessoas a fazerem formação (logo mais a não fazerem) e para maiores de 55 anos esperavam-se mais pessoas a fazer formação (logo menos a não fazer).

Para o nível de significância de 0.05 o quantil da distribuição Normal é $z_{0.975} = 1.96$. De entre todos os valores absolutos dos resíduos, os das categorias "$18 - 35$ anos" e "Mais de 55 anos" da Idade são maiores que 1.96. Portanto, as classificações correspondentes são as responsáveis pela rejeição da independência.

	Sim	Não
18-35 anos	48	7
+55 anos	17	15

Tabela 3.6: Sub-tabela da Tabela 3.4 com 2 categorias da Idade

A razão de possibilidades descreve melhor esta discrepância de valores entre faixas etárias, em relação à Formação. Para a sub-tabela 3.6 a razão de possibilidades é

$$\hat{\theta} = \frac{48 \times 15}{7 \times 17} = 6.050,$$

e substituindo os valores adequados na estatística de teste (3.10) vem que

$$\frac{\log(6.050)^2}{\frac{1}{48} + \frac{1}{15} + \frac{1}{17} + \frac{1}{7}} = 11.206.$$

Como $\chi^2_{0.95}(1) = 3.841$, a estatística anterior pertence à região crítica do teste, devendo ser rejeitada a hipótese de que $\theta = 1$. Assim assume-se que as categorias "$18 - 35$ anos" e "Mais de 55 anos" da variável Formação estão associadas, as quais, relembre-se, são as responsáveis pela dependência entre a variável Formação e Idade. Conclui-se que a possibilidade de uma pessoa fazer formação é 6 vezes maior em pessoas na faixa etária $18 - 35$ que em pessoas com mais de 55 anos.

3.4.2 Estado Civil - Formação

Também foi analisado se o estado civil poderia ter alguma influência na formação realizada. Consideraram-se duas situações possíveis, a pessoa inquirida estar de alguma forma comprometida, categoria "Casado/União de Facto", ou não estar, categoria "Solteiro/Divorciado/Viúvo".

	Sim	Não	Total
Solteiro/Divorciado/Viúvo	32	16	48
Casado/União de Facto	60	34	94
Total	92	50	142

Tabela 3.7: Tabela de contingência 2×2 das variáveis Estado Civil e Formação

O valor da estatística de Pearson correspondente à Tabela 3.7, com 1 grau de liberdade, é

$$X^2_{obs} = 0.022$$

Este valor é menor que $\chi^2_{0.95}(1) = 3.841$, logo não pertence à região de rejeição do teste perante um nível de significância de 5%, o que permite considerar a independência das variáveis. De facto, um valor de estatística de teste tão próximo de zero, como o obtido, indica desvios da independência, em termos de comparação de frequências observadas e esperadas, bastante baixos.

3.4.3 Habilitação Literária - Situação Profisional - Formação

Verifica-se agora se o percurso académico das pessoas inquiridas teve alguma influência na formação que foi ou não realizada, assim como se a situação profissional atual terá algum peso na frequência de formações.

As observações amostrais referentes à Formação e Habilitações Literárias são dadas na tabela seguinte

	Sim	Não	Total
Básico	25	34	59
Secundário	25	7	32
Superior	42	9	51
Total	92	50	142

Tabela 3.8: Tabela de contingência 3×2 das variáveis Habilitação Literária e Formação

Com um nível de significância de 5%, a estatística de teste é $X^2_{obs} = 22.3867$ e *p-value* < 0.0001, logo existem evidências estatísticas de que a independência entre as variáveis deve ser rejeitada, de acordo com o valor tabelado de $\chi^2_{0.05}(2)$. Analisando os resíduos reduzidos dados na tabela seguinte, observa-se que os resíduos reduzidos significativos pertencem

	Sim	Não
Básico	-4.715	4.715
Secundário	1.795	-1.795
Superior	3.281	-3.281

Tabela 3.9: Resíduos padronizados entre a Habilitação Literária e a Formação

às categorias ensino Básico e Secundário, sendo a primeira a que apresenta valores de resíduos reduzidos mais elevados. É possível afirmar que estas 2 categorias causam a dependência das variáveis Formação e Habilitação Literária. Logo, é possível afirmar que caso as variáveis fossem independentes, seria de esperar que houvessem mais pessoas apenas com o ensino Básico completo a frequentar formações, assim como menos pessoas com o ensino Superior.

A razão de possibilidades referente às categorias das habilitações literárias cujos resíduos reduzidos são significativos, relacionando-as com a formação (Tabela 3.10), é dado por

$$\hat{\theta} = \frac{25 \times 9}{42 \times 34} = 0.158,$$

e o valor da estatítica de teste é aqui

$$\frac{\log(0.158)^2}{\frac{1}{25} + \frac{1}{9} + \frac{1}{42} + \frac{1}{34}} = 16.712.$$

Como este valor é maior que $\chi^2(1) = 3.841$, existem evidências estatísticas para rejeitar a hipótese de que $\theta = 1$ (independência entre estas categorias), ao nível de significância de 5%. Este resultado está de acordo com a dependência concluida entre Habilitação Literária e Formação, em que estas categorias são as responsáveis. Pode-se afirmar que é mais provável uma pessoa que estudou até ao ensino superior, ou ainda estuda, frequentar formação, que alguém apenas com o ensino básico.

	Sim	Não
Básico	25	34
Superior	42	9

Tabela 3.10: Sub-tabela da Tabela 3.8 com 2 categorias da Habilitação Literária

Para a variável Situação Profissional, a sua relação com a Formação é analisada através da Tabela 3.11, cujo valor da estatística de Pearson $X^2_{obs} = 6.2795$. A tabela tem $(5-1)(2-1) = 4$ graus de liberdade. Dado que $\chi^2_{0.95}(4) = 9.4877$, o valor observado da estatística de teste não se encontra na região crítica, logo pode considerar-se que as variáveis Formação e Situação Profissional são independentes uma da outra, ao nível de signifcância de 5%.

	Sim	Não	Total
Trab. independente	20	6	26
Trab. conta outrém	38	17	55
Estudante	12	6	18
Desempregado	16	13	29
Reformado	6	8	14
Total	92	50	142

Tabela 3.11: Tabela de contingência 5×2 das variáveis Situação Profissional e Formação

3.4.4 Habilitação Literária - Situação Profisional

Intuitivamnte, pode-se considerar que existe alguma relação entre as habilitações literárias das pessoas e a sua situação profissional. Por isso procedeu-se também à análise da relação entre estas duas variáveis. Esta hipótese é considerada no sentido em que quanto mais as pessoas progridem nas suas habilitações académicas, mais provável é encontrarem-se empregadas, e vice-versa.

O teste do Qui-Quadrado correspondente à Tabela 3.12 teve como resultado $X^2_{obs} = 20.134$. O valor tabelado para comparação é $\chi^2_{0.95}(8) = 15.5073$, pois g.l.$=(3-1)(5-1) = 8$, e portanto pode dizer-se que o nível de escolaridade de uma pessoa está relacionado com a situação profissional em que se encontra.

	Trab. independente	Trab. conta de outrém	Estudante	Desempregado	Reformado	Total
Básico	11	20	1	17	10	59
Secundário	7	13	7	3	2	32
Superior	8	22	10	9	2	51
Total	26	55	18	29	14	142

Tabela 3.12: Tabela de contingência 3×5 das variáveis Habilitação Literária e Situação Profissional

Associado à Tabela 3.12 foi produzido um gráfico mosaico (Figura 3.4). Os mosaicos são bastante utilizados quando se pretende uma visualização gráfica simples e efetiva, de uma tabela de contingência, principalmente quando as suas variáveis possuem mais que duas categorias. Segundo a definição de Michael Friendly [7], este gráfico representa diretamente as contagens de uma tabela de contingência através de retângulos, cujas áreas são proporcionais à frequência da célula correspondente. Estes rectângulos são originados através de divisões horizontais e verticais alternadas do retângulo inicial. No mosaico da Figura 3.4 também estão representados os resíduos signifivativos (> 2 em valor absoluto), com um sombreamento, possibilitando a avaliação dos desvios dos dados em relação ao modelo de independência através das cores presentes.

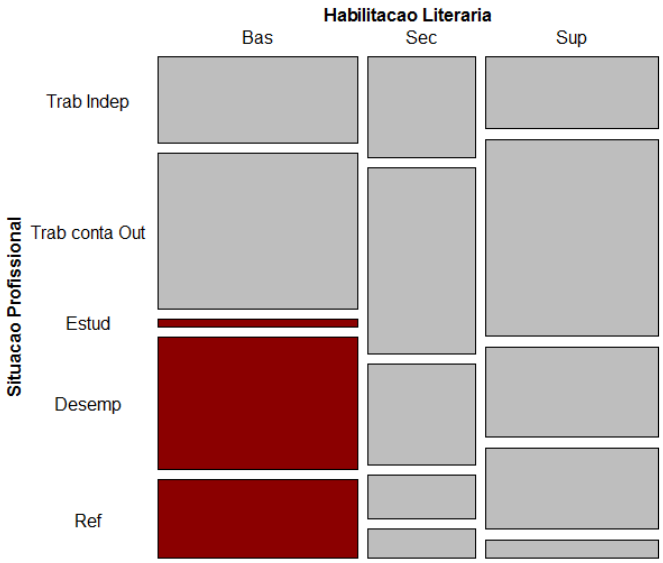

Figura 3.4: Mosaico das variáveis Habilitação Literária e a Situação Profissional

Verifica-se que existem em maior número pessoas apenas com o ensino básico, sendo estas logo seguidas do grupo de inquiridos que chegou ao ensino superior. Em relação às habilitações literárias ainda se pode ver que poucas pessoas estiveram no ensino profissional ou chegaram a um doutoramento. Relativamente à situação profissional dos inquiridos,

conclui-se que de entre as pessoas com ensino básico, a maior parte é trabalhadora por conta de outrém, apesar de o número de desempregados estar bastante próximo deste. Curiosamente, no grupo do ensino básico existem muito poucos estudantes o que pode significar que apenas estudaram até obter a escolaridade obrigatória.

Os resíduos podem aqui ser identificados através do método de sombreamento dos rectângulos baseado nos valores dos resíduos, introduzido por Friendly, sendo mais fácil detectar o padrão de afastamento da independência entre as variáveis em causa.

No mosaico da Figura 3.4 todos os resíduos reduzidos podem ser considerados não significativos, a um nível de significância de 5%, excepto nos rectângulos correspondentes aos estudantes, desempregados e reformados apenas com o ensino básico. No entanto, dois são positivos (categorias Desempregado e Reformado) e o outro é negativo (Estudante). Isto pode ser interpretado como: sob a hipótese de independência, esperar-se-ia mais pessoas com o ensino básico apenas ainda a estudar, e menos desempregadas e reformadas.

3.4.5 Formação - Região

Introduz-se agora a análise por estratos, que aqui são as regiões consideradas na fase de amostragem. Começou-se por estudar a variável Formação ao longo dos estratos. O gráfico circular 3.5 dá ênfase à diferença de valores existente entre estas duas categorias (Sim e Não).

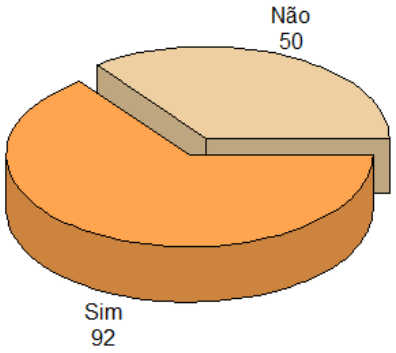

Figura 3.5: Gráfico circular relativo à variável Formação (Particulares)

Neste gráfico, é notória a discrepância existente entre o número de respostas positivas e o número de respostas negativas, sendo as respostas afirmativas quase o dobro das negativas.

Focando a atenção em cada uma das regiões, verifica-se na Tabela 3.13 que a maioria da formação que foi realizada, a nível pessoal, pertence à região do litoral centro.

Neste contexto, em que estão presentes estratos com dimensões pré-estabelecidas, deve ser feito um teste de homogeneidade em vez de um teste de independência. Isto acontece porque uma das margens da Tabela 3.13 é fixa. Com este teste vai-se tentar perceber se a formação realizada é homogénea em todos os estratos, ou seja, se o nível de formação é semelhante em qualquer região do distrito de Aveiro, ou se pelo contrário, existe alguma variação significativa entre os estratos.

	Sim	Não	Total
Litoral Norte	11	8	19
Interior Norte	18	12	30
Litoral Centro	41	21	62
Interior Centro	22	9	31
Total	92	50	142

Tabela 3.13: Tabela de contingência 4×2 da variável Formação por Região

Sob a hipótese nula de homogeneidade, obteve-se $X^2_{obs} = 1.265$ para o valor da estatística de teste do Qui-Quadrado, com $p - value = 0.7374$. Como $\chi^2_{0.95}(3) = 7.815$, aceita-se esta hipótese. Conclui-se que a distribuição de frequências da variável Formação é similar nas diferentes regiões, ou seja, existe homogeneidade na distribuição da amostra nos estratos.

Até aqui, não foi considerado o efeito que a divisão por regiões pode eventualmente exercer nas relações de duas variáveis, que foram analisadas nas secções anteriores. Nas próximas secções realiza-se um estudo por estratos destas mesmas análises, mas somente nas que envolvem a variável Formação. Isto porque os estratos foram definidos a pensar na possível diferença de comportamento da variável de interesse, em cada diferente estrato. Deste modo, serão agora examinadas tabelas de contingência com 3 fatores, em vez de tabelas de 2 fatores, com o objetivo de verificar se a independência ou dependência da variável formação com outras variáveis se mantém, quando ambas as variáveis são controladas pela região (Hipótese de Independência Condicional - Hipótese 3 da Secção 3.2). Esta hipótese é testada através do teste Cochran-Mantel-Haenszel (CMH), realizado no R através do comando *mantelhaen.test()*.

Se se decidir não rejeitar a hipótese de independência condicional, com base no *p-value* do teste, conclui-se que a formação é independente da segunda variável em causa, dados os estratos. Caso existam evidências de que esta hipótese deve ser rejeitada, será verificado se existe independência conjunta entre a variável Formação juntamente com uma outra variável e a Região (Hipótese de Independência Conjunta - Hipótese 2 da Secção 3.2). As variáveis serão denotadas pelas respetivas inicias quando forem identificadas no modelo de independência ajustado.

Dada a natureza do estudo, o significado das variáveis e a interpretação que se lhes pode dar, não faz sentido verificar a hipótese de associação homogénea (Hipótese 4 da Secção 3.2). Segundo esta hipótese, o efeito de uma variável noutra é o mesmo em cada categoria da terceira variável. No seguimento dos pares de variáveis em estudo nesta secção e adicionando a variável Região, considera-se não ter grande sentido saber, por exemplo, que em cada categoria de habilitação literária, o efeito da formação na região (ou vice-versa) é sempre o mesmo.

3.4.6 Género - Formação - Região

Verificou-se na Secção 3.4.1 que a formação realizada é independente do género da pessoa que a realiza. Pretende-se agora analisar a independência entre as variáveis quando são condicionadas pelas regiões. Para tal, foi testada a hipótese de independência nas tabelas parciais, que correspondem a cada uma das regiões, e cujas frequências observadas estão na Tabela 3.14.

		Género	
Região	Formação	Feminino	Masculino
Litoral Norte	Sim	10	1
	Não	5	3
Interior Norte	Sim	12	6
	Não	10	2
Litoral Centro	Sim	34	7
	Não	15	6
Interior Centro	Sim	18	4
	Não	3	6

Tabela 3.14: Tabela de contingência $2 \times 3 \times 4$ das variáveis Género e Formação por Região

O teste CMH forneceu um *p-value* de 0.06608, o qual indica que a hipótese nula não deve ser rejeitada, ao nível de significância de 5%. Assim, as variáveis Formação e Género sao independentes, dada a região. Relembre-se que foi assumida a independência marginal das variáveis Formação e Género. Quando estas são controladas por uma terceira variável, a Região, esta independência obviamente mantém-se. De acordo com a notação da Secção 3.2 e tomando as iniciais das variáveis para as identificar, o tipo de independência ajustada a estas três variáveis é representada por $[GR][FR]$.

3.4.7 Idade - Formação - Região

Prossegue-se com a mesma análise mas agora com a variável Idade, em substituição da variável Género, de modo a verificar se a dependência marginal verificada na Subsecção 3.4.1 entre a formação e a idade das pessoas (e ignorando a região em que se encontram) desaparece quando é adicionada uma terceira variável, ou se se mantém. Neste caso, obteve-se $p - value = 0.002$ no teste CHM aplicado à Tabela 3.15, que é um valor significativo, levando à rejeição da hipótese de independência das variáveis Formação e Idade, dada a Região, com um nível de 5%. Logo, não existe independência condicional relativamente a estas 3 variáveis e portanto a associação que se demonstrou existir entre a Formação e Idade é afetada pela Região.

		Idade		
Região	Formação	18-35 anos	36-55 anos	+55 anos
Litoral Norte	Sim	7	3	1
	Não	3	2	3
Interior Norte	Sim	8	8	2
	Não	1	8	3
Litoral Centro	Sim	25	13	3
	Não	9	5	7
Interior Centro	Sim	8	13	1
	Não	4	3	2

Tabela 3.15: Tabela de contingência $2 \times 3 \times 4$ das variáveis Idade e Formação por Região

Na Figura 3.6 encontram-se, o logaritmo das razões de possibilidades e os respetivos intervalos de confiança a 95%, para cada região, entre as categorias "$18 - 35$ anos" e "$36 - 55$ anos", e entre esta última e a categoria "Mais de 55 anos". A tabela parcial correspondente a cada região tem 3 linhas e 2 colunas, apenas $(3-1)(2-1) = 2$ razões de possibilidades locais, como as definidas na Eq. (3.11) são suficientes para se perceber em qual dos estratos está presente associação condicional entre a formação e a faixa etária.

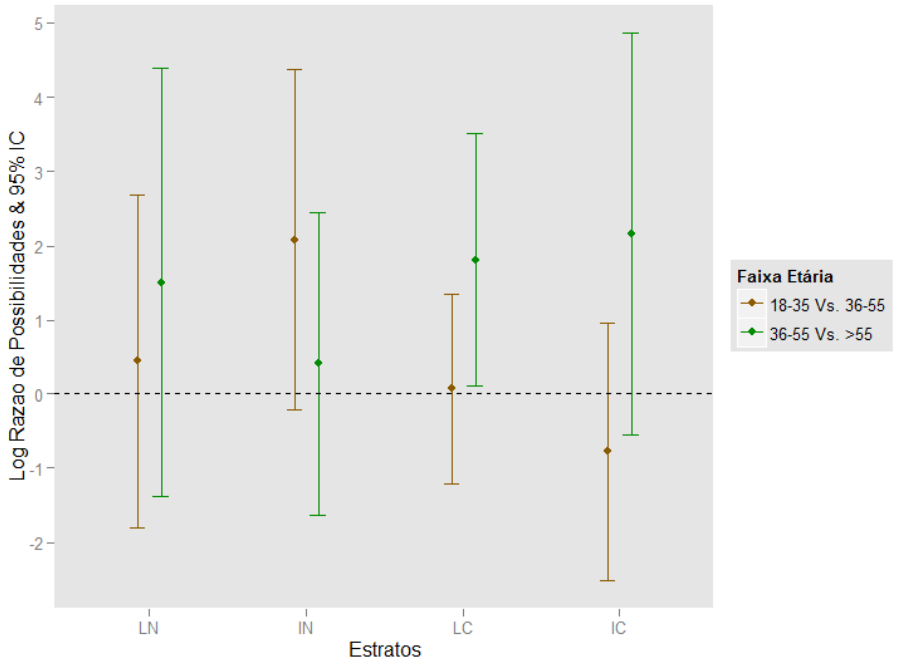

Figura 3.6: Logaritmo das Razões de Possibilidades relativas à Tabela 3.15, e IC's

Observa-se que a região Litoral Centro é a única região onde existe dependência condicional, pois nesta região, um dos intervalos do gráfico não contém o valor 0, que corresponde à independência em termos logarítmicos.

Na secção 3.4.1 concluiu-se que há dependência marginal entre a formação e a idade. Dado que a hipótese de independência condicional não se ajusta aos valores observados da Tabela 3.15, assim como hipótese de independência mútua também não (pois consequentemente as variáveis Formação e Faixa etária teriam se ser independentes), testa-se agora se a hipótese de independência conjunta destas duas variáveis com a região é adequada aos dados. Neste sentido, o teste de independência será aplicado à Tabela 3.16, considerando as combinações da Formação e Idade uma única variável, denominada Formação-Idade.

	Litoral Norte	Interior Norte	Litoral Centro	Interior Centro	Total
18-35 anos & Sim	7	8	25	8	48
36-55 anos & Sim	3	8	13	13	37
+55 anos & Sim	1	2	3	1	7
18-35 anos & Não	3	1	9	4	17
36-55 anos & Não	2	8	5	3	18
+55 anos & Não	3	3	7	2	15
Total	19	30	62	31	142

Tabela 3.16: Tabela de contingência 6×4 da variável Idade-Formação- por Região

Foi usado o teste exato de Fisher devido à existência de algumas frequências observadas muito baixas. Como este fornece um aproximado de $p - value = 0.388$ maior que a probabilidade associada à região de rejeição, a independência entre a Formação-Idade e as várias regiões não é rejeitada, ao nível de significância de 5%. Assim, as variáveis Formaçao e Idade são conjuntamente independentes dos estratos, $[IF][R]$. Isto implica independência marginal entre a Idade e a Região, e também entre a Formação e a Região.

3.4.8 Estado Civil - Formação - Idade

No estudo de uma eventual associação entre estado civil e formação, pensou-se ser mais coerente analisá-las por faixa etária e não por estratos. Isto porque o estado civil de um indivíduo é algo inerente à sua idade. Assim pretende-se averiguar se a independência das variáveis Formação e Estado civil se mantém quando estas são controladas pela Idade. Com um $p-value = 0.844$ obtido no teste de independência condicional aplicado à Tabela 3.17, decide-se que este tipo de independência é adequada neste caso. A ideia de que a idade seria uma variável de controlo, neste caso, parece ser válida. Portanto o modelo de independência é denotado por $[CI][FI]$ (C corresponde à variável Estado civil).

		Formação		
Idade	Estado Civil	Sim	Não	Total
18-35 anos	Solteiro/Divorciado/Viúvo	23	12	35
	Casado/União de Facto	25	5	30
36-55 anos	Solteiro/Divorciado/Viúvo	6	1	7
	Casado/União de Facto	31	17	48
+55 anos	Solteiro/Divorciado/Viúvo	3	3	6
	Casado/União de Facto	4	12	16

Tabela 3.17: Tabela de contingência $2 \times 2 \times 3$ das variáveis Estado Civil e Formação por Idade

3.4.9 Habilitação Literária - Formação - Região

Analisa-se de seguida se a relação de dependência demonstrada existir, na Secção 3.4.3, entre as variáveis Formação e Habilitação Literária, se altera quando são analisadas por regiões. As frequências obtidas cruzando estas duas variáveis podem ser observadas na Tabela 3.18. Como esta possui observações muito pequenas e até nulas, procurou-se estudar melhor a ligação entre estas variáveis através de gráficos apropriados para dados qualitativos, como os presentes neste estudo. Estes permitem observar as três variáveis em conjunto e demonstram de algum modo o tipo de relação que se quer conhecer.

		Habilitação Literária		
Região	Formação	Básico	Secundário	Superior
Litoral Norte	Sim	4	4	3
	Não	6	1	1
Interior Norte	Sim	4	6	8
	Não	11	1	0
Litoral Centro	Sim	8	10	23
	Não	13	3	5
Interior Centro	Sim	9	5	8
	Não	4	2	3

Tabela 3.18: Tabela de contingência $2 \times 3 \times 4$ das variáveis Habilitação Literária e Formação por Região

Inicialmente, obteve-se uma matriz de mosaicos, que permite uma rápida visualização dos dados. Esta matriz foi conseguida através da função *pairsplot* e pode ser observada na

Figura 3.7, que apresenta todos os pares possíveis de variáveis, em cada um dos mosaicos.

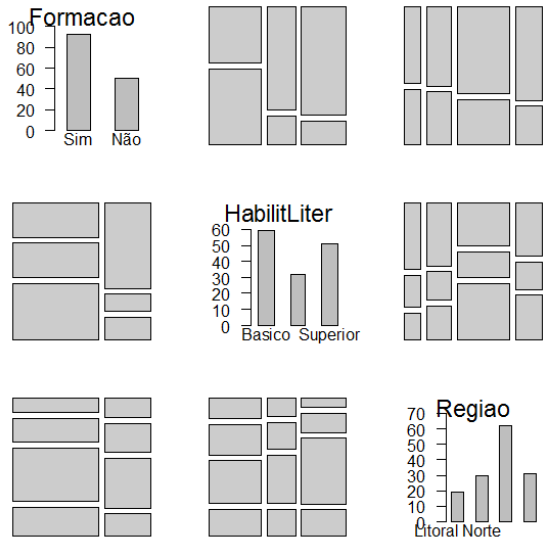

Figura 3.7: *Pairsplot* das variáveis Habilitação Literária, Formação e Região

Na primeira coluna, o segundo gráfico mostra as habilitações literárias dada a formação, onde o total de observações no ensino básico possuem a maioria das respostas Não da formação. Na análise dos resíduos, na Subsecção 3.4.3, concluiu-se que este valor deveria ser menor perante a hipótese de independência. Já o ensino superior é o que detém mais respostas Sim. Logo a formação varia consoante as habilitações literárias, o que evidencia a associação que já se concluiu existir entre estas variáveis.
Relativamente às habilitações literárias dadas as regiões (segundo gráfico da última coluna), as três categorias das habilitações apresentam proporções semelhantes em três regiões, excepto na região Litoral Centro. No entanto, a diferença de distribuição das habilitações nesta exceção, em relaçao às outras, parece ser relativamente pequena. Nos mosaicos que relacionam as variáveis Formação e Região, observa-se uma possível independência entre estas duas variáveis, conclusão já obtida na Secção 3.4.5.

Depois de analisadas as variáveis duas a duas, investiga-se a interação existente entre as três variáveis simultaneamente. Para isso utilizou-se um gráfico *doubledecker*, que é apropriado quando se tem uma variável resposta binária que se quer explicar através de outras variáveis, mostrando a dependência dessa variável dependente em relação às restantes variáveis. Nesta situação, a sua estrutura apresenta a distribuição condicional da variável Formação dadas as variáveis Região e Situação Profissional.

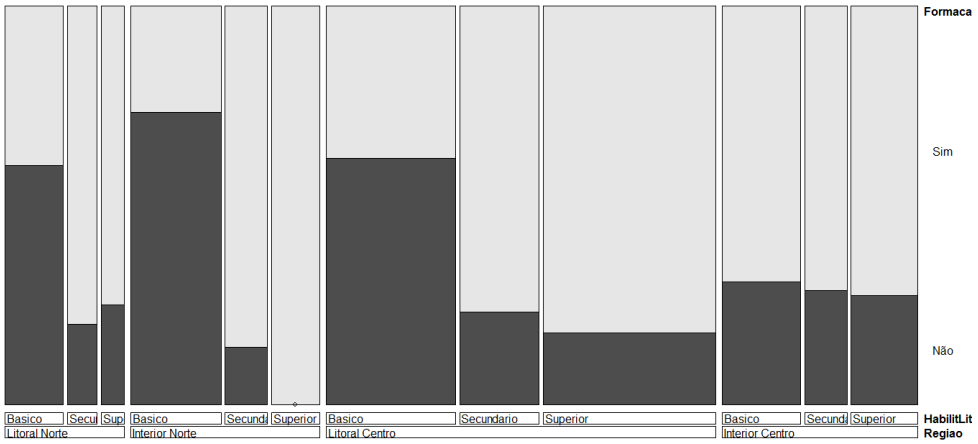

Figura 3.8: *Doubledecker* da variável Formação condicionada pela Habilitação Literária e Região

A Figura 3.8 apresenta a distribuição condicional da variável Formação dadas as variáveis Região e Habilitações Literárias. Pelo sombreamento atribuido às categorias Sim e Não, observa-se que as taxas de formação diferem entre habilitações literárias numa memsa região, assim como de região para região. No entanto, parece existir um padrão no modo como a formação varia ao longo das várias habilitações literárias, já que em todos as regiões o ensino básico apresenta um nível de formação mais baixo, tendo os ensinos secundário e superior níveis de formação mais elevados. Este padrão não é muito acentuado na região Interior Centro, onde as taxas de formação são semelhantes tanto no ensino básico, como no ensino secundário e superior. Numa primeira análise, este seria o único estrato onde se pode pensar existir independência condicional.

Contudo, para se poder fazer inferências correctas, é necessário obter gráficos que representem adequadamente o tipo de independência em análise, isto é, gráficos especializados na visualização de estruturas de independência condicional. Um exemplo destes gráficos consiste no gráfico mosaico. O sombreamento indica o grau de adequação do modelo. Porém, como os dados estão estratificados, seria incorrecto representá-los num mosaico que não tivesse em conta a variável estratificante, dado que as divisões do mosaico não são corrigidos em relação à distribuição marginal das variáveis estratificantes.

A Figura 3.9 mostra assim um painel com vários mosaicos, onde cada um permite a visualização da tabela parcial da Formação e Habilitações Literárias respetiva a uma dada Região. Todas estas subtabelas são corrigidas para a distribuição marginal da variável Região, logo todos os mosaicos têm a mesma proporção entre si. Neste painel observam-se resíduos significativos em duas regiões: Interior Norte e Litoral Centro. Deste modo, conclui-se pela rejeição da hipótese de independência condicional, sendo estes estratos os responsáveis por esta decisão. Note-se que se tinha suposto existir independência na região Interior Centro no interpretação ao gráfico 3.8. Esta indepedência é aqui verificada e constata-se que também existe no último estrato, Litoral Norte.

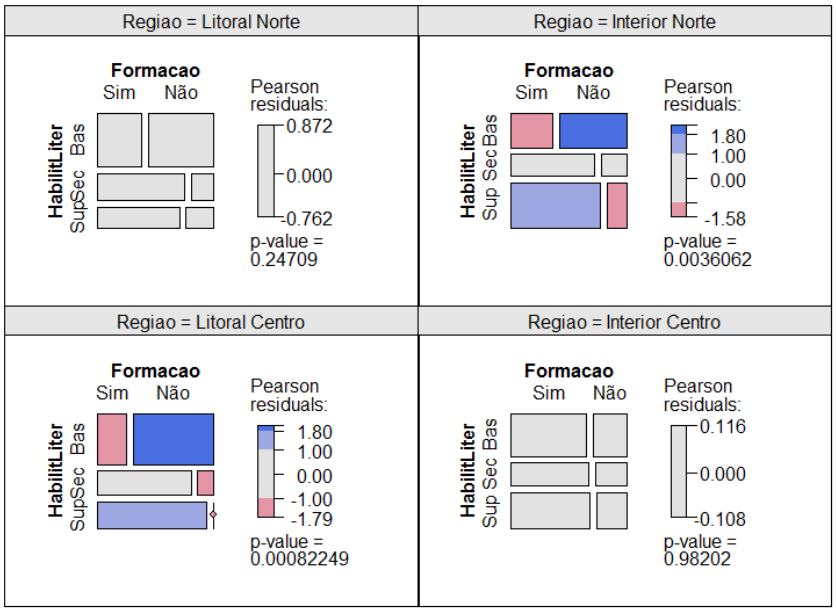

Figura 3.9: Mosaicos condicionais da Formação e Habilitação Literária dada a Região

Como consequência do não ajustamento de um modelo de independência condicional neste caso, segue-se com a verificação da possibilidade de existência de independência conjunta das habilitações literárias e formação com a região, à semelhança do que foi feito na Subsecção 3.4.7. Isto porque também neste caso a formação está associada com a habilitação literária (Subsecção 3.4.3). A junção destas duas variáveis transforma a Tabela 3.18 na seguinte, com as variáveis Formação-Habilitação e Região.

	Litoral Norte	Interior Norte	Litoral Centro	Interior Centro	Total
Básico & Sim	4	4	8	9	25
Secundário & Sim	4	6	10	5	25
Superior & Sim	3	8	23	8	42
Básico & Não	6	11	13	4	34
Secundário & Não	1	1	3	2	7
Superior & Não	1	0	5	3	9
Total	19	30	62	31	142

Tabela 3.19: Tabela de contingência da Formação - Habilitação Literária e Região

Tal como no final da Secção 3.4.7 e pelos mesmos motivos, foi aplicado o teste exato de Fisher através do qual se obteve $p - value = 0.481$, logo existem evidências para não rejeitar a hipótese de que a região é independente da formação e habilitação literária quando

estas formam uma só variável, com uma significância de 5%, e portanto o modelo adequado a este conjunto de variáveis é denotado por $[HF][R]$. As implicações entre os vários tipos de independência permitem concluir também pela independência marginal das variáveis Formação e Região, assim como das variáveis Habilitações Literárias e Região. Para completar esta análise resta descobrir a natureza da relação entre Formação e Habilitações Literárias, o que já foi feito na Secção 3.4.3 onde se constatou existir uma associação entre elas.

3.4.10 Situação Profissional - Formação - Região

Relativamente às variáveis Formação e Situação Profissional demonstrou-se serem independentes uma da outra (cf. Secção 3.4.3). Averigua-se de seguida se esta relação sofre alguma alteração quando as variáveis são divididas por regiões.

		Situação Profissional				
Região	Formação	Trab. indep.	Trab. conta outrém	Estudante	Desempregado	Reformado
Litoral Norte	Sim	4	3	1	2	1
	Não	2	2	1	0	3
Interior Norte	Sim	4	8	0	4	2
	Não	2	3	0	5	2
Litoral Centro	Sim	6	19	10	4	2
	Não	1	8	3	7	2
Interior Centro	Sim	6	8	1	6	1
	Não	1	4	2	1	1

Tabela 3.20: Tabela de contingência $2 \times 5 \times 4$ das variáveis Situação Profissional e Formação por Região

Também a Tabela 3.20 apresenta algumas observações nulas e por isso procurou-se estudar melhor a ligação entre estas 3 variáveis através de gráficos, procedendo do mesmo modo que na subsecção anterior.

À semelhança do que se fez na secção anterior, também aqui se obterão os vários gráficos onde se visualizam as 3 variáveis. Neste contexto em particular, em que se tem 5 categorias para a variável Situação Profissional, os vários gráficos facilitam ainda mais o estudo da relação entre as variáveis. Observa-se iniciaamente a matriz de mosaicos da Figura 3.10. Numa primeira observação, verifica-se que o número de respostas positivas é maior que o número de respostas negativas, tanto nas categorias da Situação Profissional como nas diferentes regiões, existindo uma única exceção nos Reformados, em que houve mais respostas negativas que positivas. Os trabalhadores por conta de outrém estão em

maior proporção em três dos quatro estratos: Interior Norte, Litoral Centro e Interior Centro, sendo estes também a categoria da situação profissional que mais respostas Sim têm na Formação. Nenhum dos submosaicos exibe simetria nem blocos alinhados com formas regulares, o que permite suspeitar da independência entre estas variáveis, havendo algumas dúvidas apenas em relação às regiões e situação profissional.

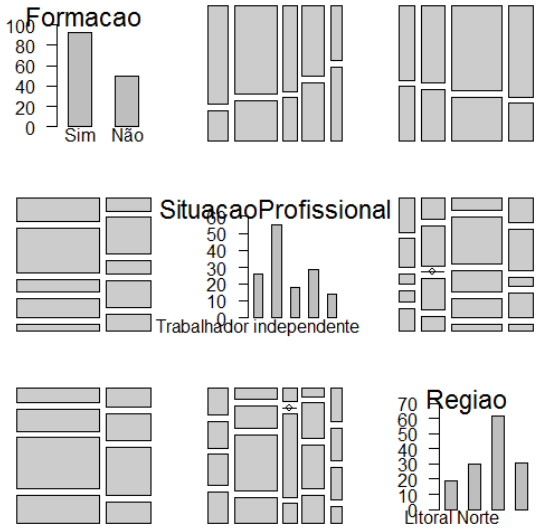

Figura 3.10: *Pairsplot* das variáveis Situação Profissional, Formação e Região

De seguida, analisam-se as interações entre as três variáveis em conjunto, usando um gráfico *doubledecker*.

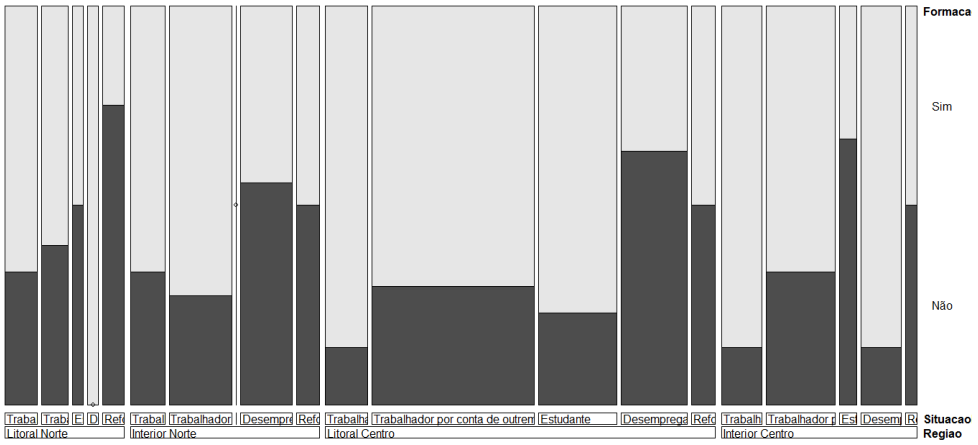

Figura 3.11: *Doubledecker* da variável Formação condicionada pela Situação Profissional e Região

Constata-se que, em cada região, existem várias oscilações da taxa de formação relativamente às várias situações profissionais. No entanto, esta distribuição desigual da taxa de formação nas categorias da Situação Profissional podem não ser significativas, dado que esta variável tem 5 categorias, e que o total de observações obtidas em cada uma delas é diferente.

Perante a presença de estratos, apresenta-se de novo um gráfico mosaico para cada estrato separadamente, que permitem visualisar as tabelas parciais da Formação e Situação Profissional, dada a Região.

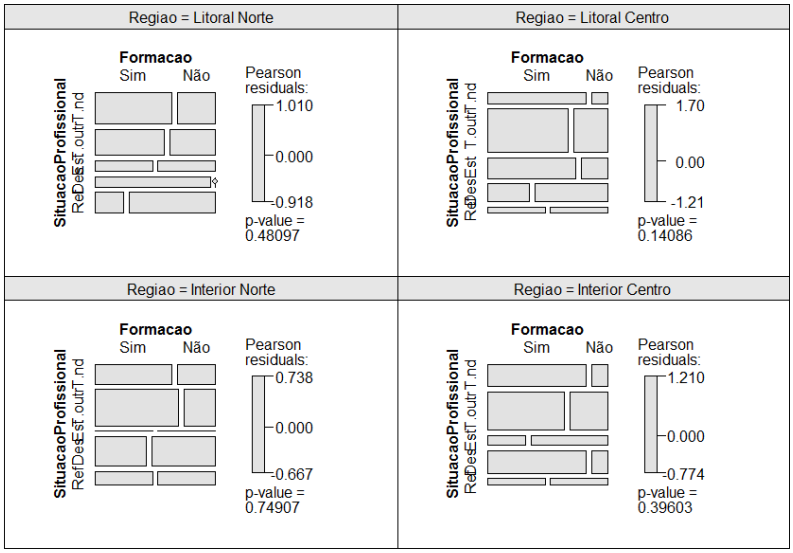

Figura 3.12: Mosaicos condicionais da Formação e Situação Profissional dada a Região

Pelas sombras apresentadas na Figura 3.12, o modelo de que a Formação e a Situação Profissional são condicionalmente independentes dada a Região não parece ser consistente com os dados pois não existem resíduos significativos. Deste modo nenhuma combinação de categorias conduz à rejeição da hipótese de independência condicional das variáveis Formação e Situação Profissional, quando condicionadas pela variável Região, e portanto pode-se adotar o modelo $[SR][FR]$. Note-se que esta independência também se verifica marginalmente.

3.4.11 Motivos da Formação

Às pessoas que responderam já terem feito alguma formação, foi-lhes perguntado qual o motivo de o terem feito. O total de cada resposta encontra-se na Tabela 3.21. Como se verifica na Tabela 3.21, a maior parte das pessoas fez as formações incentivadas pela entidade patronal (aqui considera-se mais especificamente formações de âmbito profissional) e para adquirir novos conhecimentos. A resposta "Actualização/aprofundamento de conhecimentos" também foi respondida quase tantas vezes como as citadas anteriormente.

Motivos	Total	Proporção
Incentivada pela entidade patronal	38	0.413
Actualização/aprofundamento de conhecimentos	32	0.348
Aquisição de novos conhecimentos e competências	42	0.457
Enriquecimento de curriculum	7	0.076
Iniciar nova actividade	5	0.054
Interesse/Valorização pessoal	6	0.065

Tabela 3.21: Frequências observadas e relativas dos motivos da frequência de formação nos Particulares

Já para as pessoas que não realizaram qualquer tipo de formação, os totais dos motivos apresentam-se na Tabela 3.22.

Motivos	Total	Proporção
Nunca teve interesse	44	0.88
Não têm tempo	2	0.04
Elevado custo	4	0.08

Tabela 3.22: Frequências observadas e relativas dos motivos da não frequência de formação nos Particulares

Apenas 6 pessoas, de entre 50, não referiram como motivo de não frquentarem formação o facto de nunca terem tido interesse em fazer ou em procurar alguma formação.

3.4.12 Área Profissional - Áreas de Formação

Qual a área de actividade profissional foi uma das questões do questionário, e que se pensa ser bastante relevante na frequência de formações. As áreas foram divididas usando a Classificação Nacional das Profissões de 2010 (CNP) e ao mesmo tempo, tendo em conta também as diferentes respostas obtidas, pois existem áreas na divisão da CNP para as quais nenhuma das pessoas inquiridas se inseria ou então o número de respostas nessas áreas foi muito baixo.

No gráfico 3.13 observa-se a quantidade de pessoas que fizeram ou não formação, para cada umas das áreas profissionais consideradas. Constata-se que a maioria das pessoas que responderam ao inquérito são estudantes ou domésticas, assim como pessoas empregadas em Serviços e Áreas especializadas. Também é possível reparar que em quase todas as áreas de actividade profissional, o número de respostas Sim é superior às respostas Não.

Um teste de independência foi aplicado à tabela de contingência 3.23, para verificar se a áreas de actividade profissional tem alguma influência na frequência e formação, ou não. Nesta tabela, mais de 20% das frequências esperadas são inferiores a 5, e por isso aplicou-se de novo o teste exacto de Fisher.

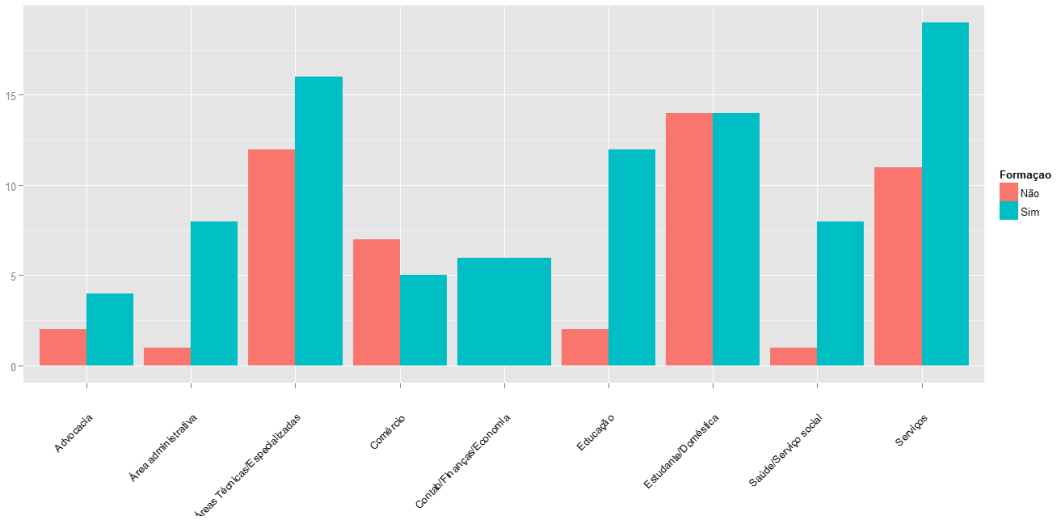

Figura 3.13: Total da variável Formação por Àrea de atividade profissional

O *p-value* de 0.031 induz à rejeiçao da hipótese nula de independência a um nível de significância de 5%, concluindo-se que estas duas variáveis estão associadas.

Área Profissional	Sim	Não	Total marginal
Advocacia	2	4	6
Área administrativa	1	8	9
Áreas Técnicas/Especializadas	12	16	28
Comércio	7	5	12
Contab/Finanças/Economia	0	6	6
Educação	2	12	14
Estudante/Doméstica	14	14	28
Saúde/Serviço social	1	8	9
Serviços	11	19	30
Total marginal	26	55	142

Tabela 3.23: Tabela de contingência 9×2 da variável Área profissional e Formação

Uma das principais questões a que se pretendia dar resposta neste estudo diz respeito às áreas em que se fazem mais formações. Tal como as áreas profissionais, também estas foram divididas, neste caso de acordo com a tabela de Áreas de Formação da Comissão Interministerial para o Emprego (CIME) (Portaria nº. 256/2005, de 16 de Março), considerando de novo apenas as secções para as quais houve respostas.

Cada pessoa podia referir várias áreas, sendo por isso esta questão de resposta múltipla. Assim cada uma das áreas de formação foi considerada como sendo uma variável dicotómica tendo por isso uma distribuição de Bernoulli, que toma o valor 1 caso a área tenha sido referida pela pessoa (sucesso) e toma o valor 0 caso contrário (insucesso). Na Tabela 3.24 pode-se observar as áreas mais referidas pelas pessoas inquiridas.

Áreas de Formação	Frequência	Proporção
Área Profissional	61	0.47
Artes e Técnicas	13	0.10
Ciências e Tecnologia	16	0.12
Comercial e Marketing	6	0.05
Línguas e Humanidades	15	0.12
Psicologia e Comportamental	8	0.06
Saúde e Cosmética	11	0.09

Tabela 3.24: Frequências observadas e relativas das áreas de formação dos Particulares

Quase metade das pessoas referiram que a formação que realizaram, ou uma das formações, foi na sua área de atividade profissional, o que é coerente com os motivos escolhidos. As seguintes áreas mais escolhidas foram Artes e Técnicas, Ciências e Tecnologia e Línguas e Humanidades, com proporções similares.

As frequências observadas que combinam estas categorias de áreas de formação com a a variável analisada antes, Àrea de atividade profissional podem ser observadas na Tabela 3.25, apenas a título de curiosidade. Isto porque não foi aplicado um teste de independência, já que cada variável apresenta bastantes categorais e a tabela tem muitas observações, o que pode comprometer a fiabilidade do teste.

	Área Profissional	Artes e Técnicas	Ciências e Tecnologia	Comercial e Marketing	Línguas e Humanidades	Psicologia e Comportamental	Saúde e Cosmética
Advocacia	3	0	0	0	0	0	1
Área administrativa	4	0	1	2	2	2	2
Áreas Técnicas/Especializadas	12	4	5	0	3	0	1
Comércio	3	1	2	0	1	1	0
Contab/Finanças/Economia	5	0	1	0	0	0	0
Educação	10	2	4	0	2	2	0
Estudante/Doméstica	3	3	2	1	4	2	2
Saúde/Serviço social	8	1	1	0	0	0	0
Serviços	13	2	0	3	3	1	5

Tabela 3.25: Tabela de contingência 9×7 das variáveis Área de formação e Área profissional nos Particulares

3.4.13 Melhoria de desempenho com a Formação

Uma conclusão interessante a retirar de entre as pessoas que realizaram formação é se esta lhes proporcionou alguma melhoria no seu desemepenho tanto a nível pessoal como a nível profissional, ou seja, se a formação foi útil de alguma forma.

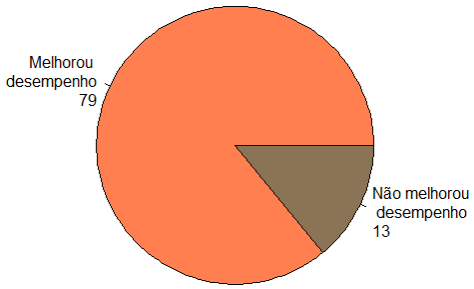

Figura 3.14: Gráfico circular relativo à melhoria de desempenho nos Particulares

De entre as 92 pessoas que fizeram formação, apenas 13 referiram não ter tirado algum proveito dela, como se pode reparar no diagrama circular 3.14.

No caso dos Particulares, perguntou-se de que modo as pessoas que responderam afirmativamente a esta questão sentiram a melhoria de desempenho, dando-se as opções de resposta da Tabela 3.26.

Consequências positivas da Formação	Total
Maior capacidade para lidar com novas tecnologias/programas informáticos	25
Adequação/actualização dos métodos de trabalho/conhecimentos	58
Eficácia no planeamento do trabalho	28
Maior capacidade de delegar funções	2
Maior capacidade de liderança	10
Maior capacidade na gestão de conflitos	15
Melhor dinâmica do grupo	12
Maior conhecimento na área	8
Maior capacidade de relacionamento interpessoal	7

Tabela 3.26: Frequências observadas das consequências positivas da Formação nos Particulares

3.4.14 Tipo e Custo da Formação e Pesquisa de Informações

As próximas questões a analisar referem-se a questões práticas relacionadas com a formação: qual o tipo de formação que as pessoas preferem (presencial ou à distância) e que preço

achariam justo ter de pagar para ter formação. Estas respostas apenas fornecem informação adicional sobre condições de formação, não tendo nenhum propósito de explicar a formação. Nas tabelas seguintes claramente existe uma tendência das pessoas para preferirem formação presencial e com o menor custo possível.

Tipo de Formação	Freq.	Freq. Relativa
Presencial	117	0.824
À distância	5	0.035
Mista	20	0.141

Tabela 3.27: Tipo de formação preferencial nos Particulares

Custo da Formação	Freq.	Freq. Relativa
10-25 €	116	0.817
25-55 €	25	0.176
55-100 €	1	0.007

Tabela 3.28: Custo justo de formação preferencial nos Particulares

Também se achou relevante saber onde a população recorre para obter informações sobre alguma formação que pretendam fazer.

Meio de Informação	Freq.	Freq. Relativa
Internet	87	0.476
Publicidade	18	0.098
Comunicação Social	15	0.082
Outros	63	0.344

Tabela 3.29: Pesquisa de Informação sobre formação nos Particulares

Cerca de metade das respostas referem a Internet como um meio de obter informações, sendo as outras opções dadas, Publicidade e Comunicação Social, as menos referidas. Existem outros meios não especificados aos quais 34% da amostra também recorre, que podem ser por exemplo referências de pessoas conhecidas.

3.4.15 Conhecimento de Entidades Formadoras

Inquiriu-se as pessoas sobre o conhecimento de centros ou entidades que dêm formação, e obteve-se a tabela de frequências para esta questão tendo em conta se a pessoa fez ou não formação. As respostas positivas estão em maior número que as respostas negativas. Parecem existir valores contraditórios na Tabela 3.30 pois existem pessoas que responderam já ter feito formação mas que não conhecem nenhuma entidade formadora. Isto pode ser explicado pelo facto de as pessoas nesta situação se terem referido a formação que tiveram a nível profissional ou por não se recordarem da entidade à qual recorrram para frequentar a formação referida.

Conhecimento / Formação	Sim	Não	Total
Sim	66	26	92
Não	20	30	50
Total	86	56	142

Tabela 3.30: Tabela de contingência 2 × 2 das variáveis Conhecimento de Entidades Formadoras e Formação nos Particulares

3.5 Análise de relações entre questões: Empresas

3.5.1 Formação - Setor de Atividade Económica

A primeira questão relevante para o estudo feita às empresas foi se ofereciam formação aos colaboradores. O gráfico circular 3.15 permite ter uma ideia geral das respostas positivas e negativas. A diferença entre a quantidade de empresas que responderam Sim e as que disseram que Não é óbvia e bastante elevada. Nos particulares as respostas positivas também foram as que prevaleceram, apesar de não ser com uma diferença tão grande.

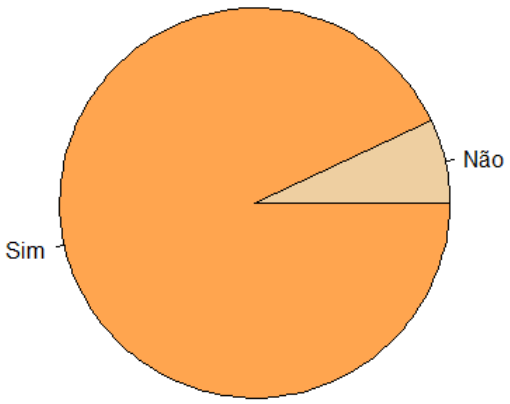

Figura 3.15: Gráfico circular relativo à variável Formação (Empresas)

Analisa-se de seguida a Formação por Setores de Atividade Económica (SAE), que para as empresas correspondem aos estratos. Na Tabela 3.31, observa-se que em 5 dos 7 setores, todas as empresas responderam Sim à questão sobre se realizam ou não formação para os seus colaboradores. Apenas nos setores Comércio por Grosso e Atividades Imobiliárias houver respostas negativas. Tal como para os Particulares, aplicou-se um teste de homogeneidade à Formação e Sectores de Atividade Económica. Foi utilizado o teste exato de Fisher, em vez do teste do Qui-Quadrado, já que existem algumas frequências esperadas inferiores a 1.

Obteve-se um *p-value* simulado de 0.00016, que é menor que o nível de significância considerado de 0.05, o que induz à rejeição da hipótese de homogeneidade entre as variáveis. Assim admite-se que a proporção de empresas que realizam formação não é a mesma em todos os setores.

Setor de Atividade Económica	Sim	Não	Total
C-Indústrias transformadoras	83	0	83
G-Comércio por grosso	75	11	86
H-Transportes e armazenagem	16	0	16
J-Atividades de informação e de comunicação	9	0	9
K-Atividades financeiras e de seguros	10	0	10
L-Atividades imobiliárias	9	5	14
M-Atividades de consultoria, científicas, técnicas e similares	12	0	12
Total	214	16	230

Tabela 3.31: Tabela de contingência 7×2 da variável Formação por Setor de Atividade Económica

Este teste não fornece os resíduos padronizados que permitem verificar onde ocorrem as fontes de dependência. No entanto, estes podem ser observados num gráfico mosaico, que também demonstra a forma como a variável Formação se distribui pelos vários SAE.

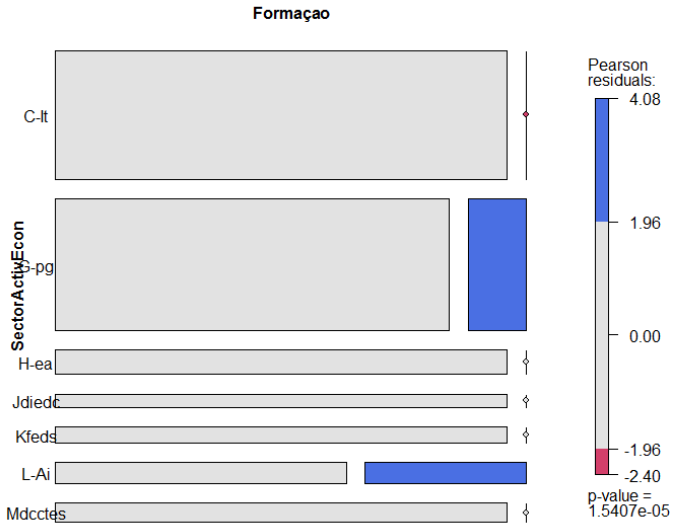

Figura 3.16: Mosaico das variáveis Formação e Setor de Atividade Económica

Tomando como referência o nível de significância 5%, existem resíduos significativos neste gráfico, indicados pelo sombreamento utilizado. De entre estes, os positivos correspondem aos dois únicos setores com empresas que responderam não disponibilizarem formação, sector Comércio por Grosso e Setor Atividades Imobiliárias. Já os resíduos negativos pertencem às Indústrias Transformadoras que não realizam formação. Em caso de homogeneidade, teriam de haver mais empresas dos setores Comércio por Grosso e Atividades Imobiliárias a terem formação e mais empresas das Indústrias Transformadoras a não o fazerem.

3.5.2 Motivos da Formação

Os motivos para se realizar formação numa dada empresa podem ser diversos, assim como também deve haver variados motivos que levam as empresas a não darem formaçao às pessoas que empregam. No questiónário foram apresentadas 7 razões para as empresas terem formação, existindo também na pergunta uma resposta aberta caso estas indicassem outros motivos que não constassem da lista de respostas. A tabela seguinte demonstra quais os motivos mais referidos pelas empresas:

Motivos	Total	Proporção
Actualização/aprofundamento de conhecimentos	144	0.673
Melhoria da imagem de empresa	9	0.042
Aumento da produtividade	38	0.178
Recebeu subsídio específico para formação	8	0.037
Aumento da motivação/satisfação dos colaboradores	23	0.108
Melhoria da qualidade do serviço	51	0.238
Diminuição do número de reclamações	11	0.051
Outros	104	0.486

Tabela 3.32: Frequências observadas e relativas dos motivos da frequência de formação nas Empresas

A maior parte das empresas referiu como motivo ou um dos motivos da formação, a necessidade de actualização e/ou aprofundamento dos conhecimentos dos funcionários. Este motivo é bastante óbvio dada a constante evolução dos métodos de trabalho e a necessidade de o mercado ser cada mais competitivo e eficiente, fazendo com que os trabalhadores necessitem de acompanhar este ritmo de desenvolvimento. Note-se que esta questão é de escolha múltipla.

Às empresas que não fazem formação foi-lhes perguntado também o motivo de tal decisão. De entre os motivos apontados, estes foram agrupados entre si pois havia várias respostas similares, formando os 3 motivos qu constam na Tabela 3.33. De entre as 16 empresas que não fazem ou fizeram formação, 3/4 delas responderam não achar necessário.

Apesar de algumas destas empresas não o terem dito, pode-se pensar que o motivo real seja a falta de financiamento para a formação, dada a dificuldade que muitas empresas atravessam hoje em dia.

Motivos	Total	Proporção
Não acham necessário	12	0.125
Não têm tempo	2	0.750
Elevado custo	4	0.125

Tabela 3.33: Frequências observadas e relativas dos motivos da não frequência de formação nas Empresas

3.5.3 Áreas de Formação

Uma questão diretamente relacionada com a realização de formação, prende-se com as áreas em que as empresas fizeram formação. Mais uma vez, devido à variedade de respostas dadas por cada empresa, foi necessário dividir as áreas referidas, resultando em 15 categorias para a variável Áreas de Formação, como se observa na Tabela 3.34.

Áreas de Formação	Total	Proporção
Ambiente	23	0.108
Ciências Informáticas	8	0.037
Comercial	59	0.276
Comportamental	34	0.159
Contabilidade e Fiscalidade	39	0.182
Finanças e Seguros	24	0.112
Higiene e Segurança	130	0.608
Informática	53	0.248
Gestão e Administração	17	0.079
Legislaçao Laboral	4	0.019
Línguas	36	0.168
Marketing	8	0.037
Qualidade	47	0.220
Serviços de Transporte	15	0.070
Técnicas	100	0.467

Tabela 3.34: Frequências observadas e relativas das áreas de formação das Empresas

Das 214 empresas da amostra que fazem formação, mais de metade (61%) referiram que

fazem formação em Higiene e Segurança no Trabalho e cerca de 47% realizam formação na sua área de atividade (Área Técnica), sendo estas as áreas mais referidas.

A distribuição das áreas de fomação por SAE pode ser examinada na Tabela 3.35. No entanto, não lhe será aplicado nenhum teste de independência, à semelhança da Tabela 3.25 e pelos mesmos motivos.

	C-Ind. Transf.	G-Com. Grosso	H-Tranp. e Armaz.	J-Act. Inf. e Comunc.	K-Act. Fin. e Seg.	L-Act. Imob.	M-Act. Cons. Cient.. Tec. e Sim.	Total
Ambiente	18	4	0	0	0	1	0	23
Ciências Informáticas	0	0	0	7	0	0	1	8
Comercial	20	31	1	1	2	4	0	59
Comportamental	23	7	2	1	0	1	0	34
Contabilidade e Fiscalidade	11	14	2	0	0	1	11	39
Finanças e Seguros	2	7	2	0	9	2	2	24
Higiene e Segurança	63	38	13	4	2	6	4	130
Informática	27	13	0	9	0	2	2	53
Gestão e Administração	4	7	1	1	0	1	3	17
Legislaçao Laboral	3	0	0	0	0	1	0	4
Línguas	23	7	3	1	0	2	0	36
Marketing	2	6	0	0	0	0	0	8
Qualidade	33	8	2	1	0	3	0	47
Serviços de Transporte	7	2	6	0	0	0	0	15
Técnicas	58	27	12	0	1	2	0	100

Tabela 3.35: Tabela de contingência 15×7 das variáveis Área de formação por Setor de atividade económica nas Empresas

3.5.4 Melhoria de desempenho com a Formação

Nas empresas em que existe formação, foi perguntado se consideravam que o desempenho dos funcionários tinha melhorado ou se, pelo contrário, nao tinham notado diferença no modo de trabalhar dos funfionários. No gráfico circular observa-se os valores obtidos nesta questão.

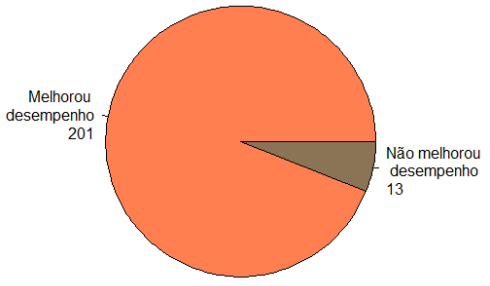

Figura 3.17: Gráfico circular relativo à melhoria de desempenho nas Empresas

Claramente, a formação que os trabalhadores tiveram foi referida como sendo algo de útil e positivo, havendo uma quantidade mínima de empresas que disseram o contrário.

3.5.5 Frequência, Tipo e Período da Formação e Pesquisa de Informações

Tal como nos particulares, também no questionário das empresas haviam questões que fornecem somente informação adicional sobre a formação realizada nas empresas de Aveiro. Estas referem-se à frequência com que se fazem formações ao longo do ano, em que período do dia a formação é feita e qual o tipo de formação preferencial.

A tabela seguinte tem a frequência com que as empresas da amostra fazem formação. Atribuiu-se para esta questão 4 categorias de frequência e acrescentou-se mais 2 categorias posteriormente devido às diferentes respostas obtidas. Esta questão foi feita apenas às empresas que responderam que faziam formação.

Frequência da Formação	Freq.	Freq. Relativa
Mensal	42	0.196
Bimestral	31	0.145
Trimestral	11	0.051
Semestral	41	0.192
Anual	55	0.257
Quando é necessário	34	0.159

Tabela 3.36: Frequência da formação realizada nas Empresas

Desta tabela conclui-se que um quarto das empresas faz formação anualmente, cerca de 16% afirmaram disponibilizar formação apenas quando acham que é necessário. Um quarto das empresas com formação realiza-a anualmente enquanto que as restantes fazem duas ou mais formações ao longo do ano.

Relativamente ao período e tipo de formação obteve-se as Tabelas 3.37 e 3.38. Mais de metade das empresas considera que a formação deve ser realizada durante o período laboral. Também se pode afirmar que quase todas preferem que seja do tipo presencial, havendo apenas uma empresa que recorre a formação à distância.

Período da Formação	Freq.	Freq. Relativa
Laboral	125	0.544
Pós-Laboral	84	0.365
Sábados	21	0.091

Tabela 3.37: Período da formação realizada nas Empresas

Tipo de Formação	Freq.	Freq. Relativa
Presencial	221	0.961
À distância	1	0.004
Mista	8	0.035

Tabela 3.38: Tipo de formação preferencial nas Empresas

Em relação ao modo como as empresas encontram informações sobre as formações que pretendam realizar, os resultados indicam que a resposta Parcerias foi a opção referida mais vezes pois aproximadamente 29% das empresas recorre a entidades formadoras com

quem já trabalharam para obter informações e requisitar novas formações. A Internet é o segundo meio de pesquisa mais utilizado, já para os Particulares este foi o meio mais escolhido.

Meio de Informação	Freq.	Freq. Relativa
Internet	62	0.200
Publicidade	37	0.120
Comunicação Social	3	0.010
Parcerias	89	0.288
Associações Comerciais	36	0.116
Entidades Formadoras	41	0.133
Entidades do próprio setor	14	0.045
Funcionários internos	12	0.039
Fornecedores	15	0.049

Tabela 3.39: Pesquisa de Informação sobre formação nas Empresas

3.5.6 Conhecimento de Entidades Formadoras

Em relação ao conhecimento das empresas inquiridads de entidades que disponibilizam formação para as empresas, obteve-se a Tabela 3.40. Na análise desta questão distingue-se entre empresas que fazem e não fazem formação, dado que as que fazem seria esperar que conhecessem algumas empresas organizadoras de Formação, ao contrário das restantes empresas.

Conhecimento / Formação	Sim	Não	Total
Sim	188	26	214
Não	4	12	16
Total	192	38	230

Tabela 3.40: Tabela de contingência 2×2 das variáveis Conhecimento de Entidades Formadoras e Formação nas Empresas

Verifica-se que são poucas as empresas que desconhecem alguma entidade formadora, comparativamente ao total da amostra. É de notar que 4 empresas que fazem formação alegaram não conhecer empresas com serviço de formação, o que pode querer dizer que têm uma entidadade formadora de referência a que recorrem usualmente, podendo estarem mais disponíveis a conhecer outras empresas de formação.

Capítulo 4

Modelos Log-Lineares

4.1 Modelos Log-Lineares para Tabelas de Contingência

Introduz-se neste capítulo a noção de Modelos Log-lineares, que pertencem à classe dos Modelos Lineares Generalizados e são indicados para observações independentes com distribuição de Poisson, como será o caso das contagens em cada célula das tabelas presentes neste trabalho. Estes modelos são bastante eficientes na análise de tabelas de contingência, no sentido em que para além de averiguarem a existência, ou não, de independência entre as variáveis, também descrevem todas as interações entre as variáveis categóricas e quantificam os efeitos que as combinações das várias categorias têm nas frequências observadas. São tendencialmente usados em tabelas com três ou mais fatores, devido ao facto de serem facilmente generalizados a estas tabelas multidimensionais. Aqui serão explicados para casos de estudo com três variáveis.

As contagens das células vão ser modeladas, e para isso, toma-sa o logaritmo das frequências esperadas das células sob independência, correspondentes às probabilidades descritas na Secção 3.2, sendo calculadas do mesmo modo que (3.4), que foram utilizadas nos testes de independência e homogeneidade da secção anterior. Estas probabilidades associadas a cada célula são expressas como uma combinação linear de parâmetros correspondentes a cada possível interação (associação) entre as variáveis da tabela. Esta transformação logarítmica permite estudar um determinado modelo multiplicativo em termos lineares. Daí a designação de Modelo Log-Linear. Como apenas se procede à modelação das contagens das células, este método não faz distinção entre variável dependente e independente.

Numa tabela de contingência tridimensional, com classificação cruzada das variáveis resposta X, Y e Z, existem vários tipos de independência (ver Secção 3.2). Os modelos log-lineares podem ser aplicados não só no caso da distribuição multinomial com probabilidades de célula π_{ijk} tais que $\sum_i \sum_j \sum_k \pi_{ijk} = 1$, mas também no caso de respostas do tipo Poisson com médias $_{ijk}$. Por exemplo, no pressuposto de independência mútua, i.e.,

$$\pi_{ijk} = \pi_{i..}\pi_{.j.}\pi_{..k}, \quad \forall i, j, k,$$

o modelo em termos dos valores esperados assume a seguinte forma log-linear:

$$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z.$$

O modelo de independência conjunta de Y em relação ao par aleatório (X, Z), traduzido por

$$\pi_{ijk} = \pi_{i\cdot k}\pi_{\cdot j\cdot}, \quad \forall i, j, k,$$

assume a forma log-linear:

$$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}.$$

O objectivo final é encontrar o modelo mais simples que melhor evidencie as relações existentes entre as variáveis subjacentes à tabela em causa. Como simples entende-se o modelo com o menor número possível de termos de associação, logo pretende-se o modelo mais parcimonioso. Os modelos variam entre si em termos das margens da tabela que ajustam e das restrições que colocam nas associações presentes nos dados. O modelo que se procura coincide, ou esté entre, o modelo de independência mútua e o modelo saturado.

O modelo saturado descreve a estrutura mais complexa que é possível ajustar aos dados e tem tantos parâmetros quanto células na tabela de contingência. Deste modo, fornece um ajustamento perfeito aos dados, ou seja, descreve perfeitamente as frequências observadas n_{ijk} da tabela, que têm distribuição de Poisson com valores esperados m_{ijk}. Como as fórmulas dos modelos log-lineares usam o valor esperado correspondente a cada probabilidade conjunta das células da tabela, com distribuição multinomial, também podem ser aplicadas as contagens de Poisson. O seguinte modelo saturado permite todas as interacções possíveis entre os três fatores:

$$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^K + \lambda_{ij}^{XY} + \lambda_{ik}^{XK} + \lambda_{jk}^{YK} + \lambda_{ijk}^{XYK} \tag{4.1}$$

Nesta equação as frequências esperadas m_{ijk} são função de uma constante que representa o efeito da média global (λ), de termos de efeitos principais ($\lambda_i^X, \lambda_j^Y, \lambda_k^K$) e de termos de efeitos de interação dupla e tripla entre as variáveis ($\lambda_{ij}^{XY}, \lambda_{ik}^{XK}, \lambda_{jk}^{YK}, \lambda_{ijk}^{XYK}$). O número de parâmetros de cada termo depende do efeito que ele representa (principal ou interação) e das categorias das variáveis consideradas nesse termo. Já o modelo de independência possui na combinação linear apenas a constante e os termos de efeitos principais, pois não existe interação entre os fatores. O modelo mais parcimonioso que se pretende alcançar isola em si apenas os efeitos significativos, definindo os restantes efeitos a zero.

Geralmente estes modelos são hierárquicos, no sentido em que modelos de maior ordem englobam modelos de menor ordem (os termos e respetivos parâmetros do modelo com menor ordem estão obrigatoriamente presentes no modelo de maior ordem). Devido a esta propriedade, a ideia base de qualquer método de escolha de um modelo consiste em testar modelos reduzidos contra modelos maiores, caso não exista uma hipótese para o modelo *a piori*. O processo mais utilizado para a verificação de todos os modelos parte do modelo saturado e elimina sucessivamente as interações de maior ordem. Em cada eliminação é obtido um modelo com menos termos para o qual se calcula a estatística de teste.

Ajustamento de Modelos

O primeiro passo no processo de ajustamento de um determinado modelo é o cálculo das frequências esperadas, através de fórmulas apropriadas que dependem do modelo log-linear que está a ser considerado. Com estes valores, testa-se a hipótese nula de que o modelo log-linear fornece um bom ajustamento aos dados. A estatística de teste utilizada é a razão de verossimilhança (ou estatística de desvio)

$$G^2 = 2\sum_{i,j,k} n_{ijk}\, log\left(\frac{n_{ijk}}{\hat{m}_{ijk}}\right)$$

que tem distribuição assintótica χ^2, com o número de graus de liberdade dado pelo total das células da tabela menos o número de parâmetros estimados no modelo. A hipótese nula aqui equivale a testar se os termos que não figuram num modelo são de facto nulos, para que possam ser eliminados, obtendo-se em cada passo um modelo mais simples.

Comparação de Modelos

De entre os modelos obtidos na fase anterior, é necessário escolher de entre os modelos aceites, qual deles providencia o melhor ajustamento dos dados. Considera-se que um modelo M_1 está "encaixado"no modelo M_2 se o modelo M_1 contém apenas alguns dos parâmetros de M_2. Devido à sua propriedade de aditividade, a estatística G^2 é a única que se pode usar para comparar modelos encaixados, através da diferença dos seus respetivos valores G^2.

Sendo M_1 o modelo com menor número de parâmetros, pode-se comparar o grau de ajustamento dos dois modelos através da estatística $G^2(M_1|M_2) = G^2(M_1) - G^2(M_2)$, com $df(M_1) - df(M_2)$ graus de liberdade (número de parâmetros eliminados). Com esta expressão testa-se se os termos que não são comuns aos dois modelos, não são de facto importantes no modelo M_1, sendo portanto nulos. Caso não se rejeite esta hipótese, o modelo mais pequeno (encaixado) é suficiente para explicar as frequências observadas.

Uma vez obtido o modelo que melhor se ajusta à tabela em estudo, será possível atribuir uma estrutura aos dados e identificar as relações entre as variáveis. Adicionalmente e numa análise mais profunda, também é possível quantificar os efeitos que as variáveis e as interações entre elas exerceram sobre os dados e analisar a sua significância, através dos estimadores dos seus parâmetros.

Qualquer modelo log-linear definido para um determinado conjunto de dados, está associado a um tipo de independência entre 3 fatores, descritos na Secção 3.2.

1. **Independência Mútua [X][Y][Z]**

$$\log\, \hat{m}_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z \qquad (M1)$$

2. **Independência Conjunta/Parcial [XY][Z]**

$$\log\, \hat{m}_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_k^Z \qquad (M2)$$

3. **Independência Condicional [XZ][YZ]**

$$\log \hat{m}_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} \qquad (M3)$$

4. **Associação Homogénea [XY][XZ][YZ]**

$$\log \hat{m}_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_k^{XZ} + \lambda_{jk}^{YZ} \qquad (M4)$$

A modelação log-linear será aplicada aos conjuntos de variáveis analisados na Secção 3.4. Apesar de todas as vantagens e conclusões que a análise de um modelo log-linear pode providenciar, apenas se pretende aqui confirmar se o modelo que se encontrou nas análises com três fatores, realizadas na Secção 3.4, é o mais adequado, e não calcular parâmetros com fins de interpretação dos efeitos das variáveis, tanto indivualmente como em conjunto. Um detalhe importante é que os estratos são incluidos como fatores/variáveis na tabela e por isso a amostragem estratificada não causa problemas numa análise através do modelo log-linear.

No programa estatístico usado, R, existem várias funções à escolha nesta matéria. Optou-se por usar a função *loglm()* para realizar os cálculos necessários à obtenção dos valores G^2 (desvios).

4.1.1 Particulares: Género - Formação - Região

O primeiro conjunto de variáveis ao qual se aplicou a modelação log-linear engloba as variáveis da Subsecção 3.4.6: Formação, Género e Região, para as quais foi identificada independência condicional $[GR][FR]$. Os desvios G^2 dos vários modelos possíveis de aplicar a um conjunto de 3 fatores apresentam-se na Tabela 4.1. Estes desvios são sempre em relação ao modelo saturado, que é sempre um modelo que se ajusta perfeitamente aos dados.

Modelos	G^2	g.l.	*p-value*
$[G]\,[F]\,[R]$	13.921	10	0.177
$[G]\,[FR]$	12.656	7	0.081
$[F]\,[GR]$	12.341	7	0.090
$[R]\,[GF]$	10.384	9	0.320
$[FG]\,[RG]$	8.804	6	0.185
$[GF]\,[RF]$	9.118	6	0.167
$[GR]\,[FR]$	11.075	4	<0.05
$[GF]\,[GR]\,[FR]$	7.239	3	0.065
$[GFR]$	0.000	0	1

Tabela 4.1: Desvios, graus de liberdade e *p-values* dos modelos ajustados às variáveis Género, Formação e Região

Todos as probabilidades são significativas ao nível de 5%, exceto no modelo $[GR][FR]$, o que significa que estes fornecem um bom ajustamento aos dados. Logo será necessário recorrer à comparação de modelos, o que permitirá descobrir que parâmetros destes modelos podem ser considerados nulos, sem prejuízo no grau de ajustamento. Como se pretende o modelo mais económico em termos de parâmetros, vai-se testar o modelo $[G][F][R]$ contra os outros modelos aceites como bons ajustamentos, de um modo hierárquico.

Através da Tabela 4.2, chega-se à conclusão de que o modelo de independência mútua $[G][F][R]$ é um bom ajustamento comparativamente com os restantes modelos, em termos de significância a 5%. Isto significa que todos os termos de interação entre variáveis, não presentes no modelo $[G][F][R]$, não são necessários no sentido em que um modelo sem eles se ajusta tão bem como com eles e podem ser considerados nulos. Assim o modelo de independência mútua $[G][F][R]$ é o modelo com um menor número de parâmetros que melhor se ajusta aos dados. Daqui conclui-se que as variáveis Género, Idade e Região são todas independentes entre si. Recorde-se que no capítulo anterior foi obtido o modelo $[GR][FR]$. Esta conclusão está correta pois este tipo de independência é um caso particular da independência mútua (tal como todos os outros tipos).

Modelos em comparação			G^2	g.l.	*p-value*
$[GF]\,[GR]\,[FR]$	*vs.*	$[FG]\,[RG]$	1.565	3	0.667
		$[GF]\,[RF]$	1.879	3	0.598
$[FG]\,[RG]$	*vs.*	$[F]\,[GR]$	3.538	1	0.060
		$[R]\,[GF]$	1.580	3	0.664
$[GF]\,[RF]$	*vs.*	$[G]\,[FR]$	3.538	1	0.060
		$[R]\,[GF]$	1.266	3	0.737
$[G]\,[F]\,[R]$	*vs.*	$[G]\,[FR]$	1.266	3	0.737
		$[F]\,[GR]$	1.580	3	0.664
		$[R]\,[GF]$	3.538	1	0.060

Tabela 4.2: Comparação entre os modelos aceites da Tabela 4.1

4.1.2 Particulares: Idade - Formação - Região

Na respetiva subsecção em que se testam as variáveis Idade, Formação e Região (Subsecção 3.4.7), foi-lhes associada independência conjunta através do modelo $[R][IF]$. Verificando os restantes modelos que podem ser ajustados, tem-se os desvios G^2 na Tabela 4.3, onde se observa que vários outros modelos podem ser considerados adequados, a um nível de significância de 5%, na descrição da relação entre estas três variáveis: $[F][IR]$, $[R][IF]$, $[FI][RI]$, $[IF][RF]$ e $[IF][IR][FR]$.

Modelos	G^2	g.l.	*p-value*
$[I]\,[F]\,[R]$	28.077	17	<0.05
$[I]\,[FR]$	26.811	14	<0.05
$[F]\,[IR]$	18.283	11	0.075
$[R]\,[IF]$	15.606	15	0.409
$[FI]\,[RI]$	5.813	9	0.759
$[IF]\,[RF]$	14.340	12	0.280
$[IR]\,[FR]$	17.018	8	<0.05
$[IF]\,[IR]\,[FR]$	5.044	6	0.538
$[IFR]$	0.000	0	1

Tabela 4.3: Desvios, graus de liberdade e *p-values* dos modelos ajustados às variáveis Idade, Formação e Região

Prossegue-se assim com novas comparações de modelos, agora entre os modelos aceites como bons e com menos parâmetros, e todos os restantes. Nas duas primeiras comparações na Tabela 4.4 aceita-se os modelos $[FI][RI]$ e $[IF][RF]$, em detrimento do modelo de associação homogénea, o que indica que os termos $[IR]$ e $[FR]$ correspondem a interações não presentes aqui e por isso não pertencerão ao modelo final, sendo considerados nulos. Daqui já se poderia afirmar que um dos dois modelos de independência conjunta com *p-value* não significativo na Tabela 4.3 será o modelo encontrado para descrever a relação entre estas três variáveis. Isto pode ser observado nas últimas três comparações da Tabela 4.4, onde o modelo de independência conjunta $[R][IF]$ é o que melhor se ajusta aos dados e melhor modela as frequências observadas da tabela de contingência que cruza estas variáveis. Esta conclusão está de acordo com a obtida em 3.4.7.

Modelos em comparação			G^2	g.l.	*p-value*
$[IF]\,[IR]\,[FR]$	*vs.*	$[FI]\,[RI]$	9.297	6	0.158
	vs.	$[IF]\,[RF]$	0.769	3	0.857
$[FI]\,[RI]$	*vs.*	$[F]\,[IR]$	12.470	2	0.001
	vs.	$[R]\,[IF]$	9.793	6	0.134
$[IF]\,[RF]$	*vs.*	$[R]\,[IF]$	1.266	3	0.737

Tabela 4.4: Comparação entre os modelos aceites da Tabela 4.3

4.1.3 Particulares: Estado Civil - Formação - Idade

Em relação às variáveis Estado Civil e Formação supôs-se existir um condicionamento destas pela Idade, um presuposto que se verificou ser válido na Secção 3.4.8. Analisa-se agora se este é o modelo mais parsimonioso para este conjunto de variáveis ou se outro com menor número de parâmetros também será adequado.

Modelos	G^2	g.l.	*p-value*
$[C]\,[F]\,[I]$	41.999	7	<0.01
$[C]\,[FI]$	29.528	8	<0.01
$[F]\,[CI]$	17.753	5	<0.01
$[I]\,[CF]$	41.886	6	<0.01
$[FC]\,[IC]$	17.640	4	<0.01
$[CF]\,[IF]$	29.415	4	<0.01
$[CI]\,[FI]$	5.282	3	0.152
$[CF]\,[CI]\,[FI]$	5.243	2	0.073
$[CFI]$	0.000	0	1

Tabela 4.5: Desvios, graus de liberdade e *p-values* dos modelos ajustados às variáveis Estado Civil, Formação e Idade

Apenas os modelos $[CI][FI]$ e $[CF][CI][FI]$ podem ser considerados. Na sua comparação na Tabela 4.6, tem-se $p - value = 0.842 > 0.05$. Logo aceita-se a hipótese de que o parâmetro $[CF]$ é nulo e portanto o modelo de independência condicional do estado civil e formação dada a faixa etária, $[CI][FI]$, dá um bom ajustamento aos dados, tal como se tinha constatado anteriormente.

Modelos em comparação	G^2	g.l.	*p-value*
$[CI]\,[FI]$ *vs.* $[CF]\,[CI]\,[FI]$	0.040	1	0.842

Tabela 4.6: Comparação entre os modelos aceites da Tabela 4.5

4.1.4 Particulares: Habilitação Literária - Formação - Região

Na Secção 3.4.9 concluiu-se que existe neste caso dependência condicional. Pretende-se agora encontrar um modelo de modo a verificar isto.

Os desvios dos vários modelos estão registados na Tabela 4.7. Através dos valores de *p-value* aqui observados, apenas quatro destes modelos se ajustam bem aos dados: $[FH][FR][HR]$, $[FH][FR]$, $[FH][HR]$, $[R][FH]$.

Recorrendo de novo ao processo de comparação de modelos, começa-se por testar o modelo $[R][FH]$ contra os outros três modelos, obtendo-se os resultados da Tabela 4.8. Nenhum dos *p-values* da tabela é significativo, logo aceita-se a hipótese que os termos que estão nos modelos maiores e não estão no modelo menor $[R][FH]$ podem ser considerados nulos. Este modelo é o modelo mais económico em termos de parâmetros que fornece um bom ajustamento aos dados. Assim, pode-se considerar que existe independência conjunta dos factores Formação e Habilitações Literárias, face ao factor Região. Esta conclusão está de acordo com a análise realizada em 3.4.9.

Modelos	G^2	g.l.	*p-value*
$[F]\,[H]\,[R]$	38.253	17	<0.01
$[F]\,[HR]$	32.595	11	<0.01
$[H]\,[FR]$	36.988	14	<0.01
$[R]\,[FH]$	15.575	15	0.411
$[HF]\,[RF]$	14.310	12	0.281
$[FH]\,[RH]$	9.917	9	0.357
$[FR]\,[HR]$	31.329	8	<0.01
$[FH]\,[FR]\,[HR]$	9.067	6	0.170
$[FHR]$	0.000	0	1

Tabela 4.7: Desvios, graus de liberdade e *p-values* dos modelos ajustados às variáveis Habilitação Literária, Formação e Região

Modelos em comparação			G^2	g.l.	*p-value*
$[R]\,[FH]$	*vs.*	$[HF]\,[RF]$	1.265	3	0.738
$[R]\,[FH]$	*vs.*	$[FH]\,[RH]$	5.658	3	0.130
$[R]\,[FH]$	*vs.*	$[FH]\,[FR]\,[HR]$	6.508	9	0.688

Tabela 4.8: Comparação entre os modelos aceites da Tabela 4.7

A título de curiosidade, note-se que bastava a terceira comparação da Tabela 4.8 para se assumir o modelo $[R][HF]$ como o melhor modelo. Isto porque, como existem evidências estatísticas em como o modelo deve ser aceite, verificava-se de imediato que os termos $[FR]$ e $[HR]$ podem ser tomados como nulos, e consequentemente os modelos $[FH][FR]$ e $[FH][HR]$ podiam ser rejeitados.

4.1.5 Particulares: Situação Profissional - Formação - Região

Por último, averigua-se se a independência condicional existente entre as variáveis Formação e Situação Profissional, quando controladas pela Região, é mesmo o modelo mais económico em termos de parâmetros para descrever a relação entre estas três variáveis ou se outro existe, mas que não foi averiguado na secção anterior (pois o principal objetivo era verificar a presença de independência condicional).

Examinando os desvios G^2 da Tabela 4.9, percebe-se que todos os modelos nela presentes podem dar bons ajustamentos dos dados. Assim, tal como na primeira análise efetuada nesta secção, a comparação de modelos é feita entre o modelo de independência mútua e todos os outros modelos possívis de ajustar a este conjunto de variáveis.

Modelos	G^2	g.l.	*p-value*
$[F]\,[S]\,[R]$	39.953	31	0.130
$[F]\,[SR]$	18.245	19	0.506
$[S]\,[FR]$	38.688	28	0.086
$[R]\,[FS]$	33.750	27	0.174
$[SF]\,[RF]$	32.484	24	0.115
$[FS]\,[RS]$	12.040	15	0.676
$[FR]\,[SR]$	16.979	16	0.387
$[FS]\,[FR]\,[SR]$	11.113	12	0.519
$[FSR]$	0.000	0	1

Tabela 4.9: Desvios, graus de liberdade e *p-values* dos modelos ajustados às variáveis Situação Profissional, Formação e Região

De entre as três primeiras comparações da Tabela 4.10, conclui-se que o modelo $[SF][RF]$ não se ajusta bem aos dados. Logo o termo $[SR]$ deve constar no modelo que for ajustado. Nos outros dois modelos de independência condicional que não têm *p-values* significativos na comparação com o modelo de associação homogénea, figura este termo. Comparando estes dois modelos com os que representam independência conjunta, e tendo em conta as condições em que se podem comparar modelos, verifica-se que apenas o modelo $[F][SR]$ deve continuar a ser verificado, sendo os outros dois ignorados. Isto não surpreende dado que nas primeiras três comparações observa-se que os parâmetros $[SF]$ e $[FR]$ podem ser tomados como nulos. Para finalizar, o modelo de independência não deve ser escolhido pois a sua comparação com o de independência conjunta na última linha tem um valor de significância significativo. Assim, o modelo com menos parâmetro e que simultaneamente dá um bom ajustamento neste caso é o modelo em que a situação profissional e a região são conjuntamente independentes da formação, ou em notação, $[F][SR]$.

Modelos em comparação			G^2	g.l.	*p-value*
$[SF]\,[SR]\,[FR]$	*vs.*	$[FS]\,[RS]$	0.927	3	0.819
	vs.	$[SF]\,[RF]$	21.370	12	0.045
	vs.	$[FR]\,[SR]$	5.866	4	0.209
$[FS]\,[RS]$	*vs.*	$[F]\,[SR]$	6.205	4	0.184
	vs.	$[R]\,[SF]$	21.709	12	0.041
$[FR]\,[SR]$	*vs.*	$[S]\,[FR]$	21.709	12	0.041
	vs.	$[F]\,[SR]$	1.266	3	0.737
$[F]\,[SR]$	*vs.*	$[F]\,[S]\,[R]$	21.709	12	0.041

Tabela 4.10: Comparação entre os modelos aceites da Tabela 4.9

Contudo este modelo não faz muito sentido em termos de interpretação das variáveis. Pode-se continuar a considerar o modelo de independência condicional $[FR][SR]$, encontrado na Subsecção 3.4.10, já que este é um dos modelos aceitáveis imediatamente antes do encontrado.

Capítulo 5

Conclusões

Ao longo deste trabalho procurou-se explorar várias questões associadas à formação realizada no Distrito de Aveiro, em particulares e em empresas. O primeiro objetivo traçado foi o de encontrar uma estimativa da proporção de pessoas que frequentaram formações pelo menos uma vez, para os particulares, e uma estimativa das empresas que disponibilizam formação aos colaboradores. Numa segunda parte foram estudadas algumas variáveis que poderiam condicionar a frequência de formação, tanto ao nível de características inerentes ao indivíduo, por exemplo o género e idade, como em aspetos da vida do mesmo, tal como o estado civil, a habilitação literária e a situação profissional. Em relação às empresas foi analisada a formação só por Setor de Atividade Económica. Outras questões que se colocaram às duas populações em simultâneo, referiam-se à área da formação praticada e as condições de realização de formação, como por exemplo o preço, o período do dia, o acesso a informação, entre outros.

Foi escolhido o processo de amostragem aleatória estratificada como o processo a ser utilizado e definidos os estratos para as duas populações: a localidade/região nos particulares e o SAE nas empresas. Depois de obtida a dimensão da amostra, em particulares e empresas, e o número de elementos de cada estrato que nele deviam constar, calcularam-se as proporções estimadas da proporção de elementos que frequentaram formação, assim como os respetivos intervalos de confiança, recorrendo a fórmulas aqui demonstradas. Concluiu-se que a proporção de formação realizada é, em ambas as populações, maior que 50%, chegando nas empresas quase a totalidade destas.

Através do teste de independência do Qui-Quadrado, observou-se existir dependência da variável Formação com as variáveis Idade e Habilitação Literária, e a independência em relação às variáveis Género, Estado Civil e Situação Profissional. Daqui conclui-se que a faixa etária em que uma pessoa residente em Aveiro se encontra, assim como o seu grau académico à data do inquérito, têm influência na frequência de formação. Já o género da pessoa, o seu estado civil e a sua ocupação atual nada têm haver com a frequência de formações. Também foi avaliada a existência de associação ou não entre a habilitação académica que uma pessoa detém com a sua situação profissional, verificando-se que existe

associação entre elas. Ainda neste contexto bivariado, o teste de homogeneidade do Qui-Quadrado indicou haver homogeneidade da formação em relação às regiões, ou seja, as proporções de formação realizada e não realizada distribuem-se de forma semelhante em todos as 4 regiões/estratos.

Com a introdução de uma terceira variável, a Região, aos pares anteriores que incluem a Formação, foi testada a hipótese de independência condicional dos pares, dada a Região. Concluiu-se que a independência marginal constatada entre duas variáveis se mantém quando são controladas pela região. Nos pares em que existe dependência marginal, Idade-Formação e Habilitação Literária-Formação, não se verificou esta independência condicional. Para estes, foi averiguada a hipótese de independência conjunta da região, que se demonstrou ser válida. O par Estado Civil-Formação foi o único caso em que a variável condicionante foi a Idade em vez da Região, e também para estes a independência é condicional.

Relativamente às empresas, o teste de homogeneidade do Qui-Quadrado forneceu evidências de que não existe homogeneidade, logo em alguns setores/estratos a distribuição das categorias Sim e Não da variável Formação não é similar à distribuição noutros estratos.

Em complemento ao estudo, foram examinadas outras variáveis, com um interesse mais prático sobre a organização e as condições da formação, assim como as áreas em que se fazem formações.

Os principais motivos referidos por um particular que já tenha feito formação prende-se com motivos profissionais, e também com o objetivo de atualizar conhecimentos e adquirir novos. Este último foi o motivo mais referido pelas empresas. A maioria das pessoas que nunca fizeram formação referiu nunca ter tido interesse em fazê-lo e a maioria das empresas na mesma situação não considera necessário.

A área profissional de um indivíduo está associada com a formação feita e a área de formação mais referida é a área profissional, seguida pelas áreas Ciência e Tecnologia e Línguas e Humanidades. Já nas empresas as áreas com mais interesse são Higiene e Segurança e áreas técnicas. Grande parte de particulares e empresas inquiridas disse ter havido melhorias de desempenho com a formação.

A formação presencial é a mais referida pelas duas populações. Os particulares admitem preferir fomação a menor custo e quase todos recorrem à internet para obter informações. Já nas empresas, recorrem tanto à internet (e em menor proporção) como a parcerias que já possuem com determinadas entidades. A frequência de realização de formação varia muito entre empresas e o horário mais usado é o laboral. O conhecimento de entidades que organizam formação é maior que o não conhecimento em ambas as populações.

Por fim, os Modelos Log-Lineares aplicados às análises entre três fatores permitiram ajustar um modelo aos dados, reforçando as escolhas do tipo de independência atribuidas em cada caso, ou dando outras, também adequadas mas com menor número de parâmetros de associação. Os resultados aqui obtidos confirmam os do capítulo anterior, com exceção nos conjuntos de variáveis Género-Formação-Região e Situação Profissional-Formação-Região. No que diz respeito ao primeiro conjunto, foi ajustado um modelo de

independência mútua. Quanto ao segundo o modelo encontrado foi o de independência conjunta da Formação com a Situação Profissional e Região, notando que este não tem muito sentido dado o significado das variáveis.

As conclusões aqui conseguidas podem ter utilidade por parte de empresas que organizem formações, tendo aqui acesso tanto a caraterísticas das pessoas que de algum modo influenciam a sua decisão de fazerem formação, como quais as condições preferidas em que as pretendem fazer. Esta última observação também tem fundamento relativamente a empresas.

Bibliografia

[1] Alan Agresti. *Categorical data analysis*. Wiley-Interscience, 2002.

[2] Alan Agresti. *An introduction to categorical data analysis*. Wiley-Interscience, 2007.

[3] Razia Azen and Cindy M Walker. *Categorical data analysis for the behavioral and social sciences*. Routledge, 2010.

[4] Ronald Christensen and R Christensen. Log-linear models and logistic regression. 1997.

[5] William G. Cochran. *Sampling techniques*. John Wiley & Sons, Inc, 1977.

[6] Stephen E Fienberg. *The analysis of cross-classified categorical data*. Springer, 2007.

[7] Michael Friendly. *Visualizing categorical data*. Sas Inst, 2000.

[8] S. J. Haberman. *The Analysis of Residuals in Cross-Classified Tables*. Biometrics 29, 1973.

[9] David C Howell. *Fundamental Statistics for the Behavioral Sciences*. Cengage Learning, 2010.

[10] David C Howell. *Statistical methods for psychology*. Cengage Learning, 2011.

[11] Paul S Levy and Stanley Lemeshow. *Sampling of populations: methods and applications*. John Wiley & Sons, Inc, 2012.

[12] Sharon L Lohr. *Sampling: design and analysis*. Cengage Learning, 2010.

[13] William Navidi. *Probabilidade e Estatística para Ciências Exatas*. McGraw Hill Brasil, 1998.

[14] R. Lyman Ott. *An Introduction to Statistical Methods and Data Analysis*. Brooks/Cole, Cengage Learning, 2010.

[15] G Gerald Peter Quinn and Michael J Keough. *Experimental design and data analysis for biologists*. Cambridge University Press, 2002.

[16] R.

[17] Poduri SRS Rao. *Sampling methodologies*. Chapman & Hall/CRC, 2000.

[18] Elisabete Reis and Raul Moreira. *Pesquisa de mercado*. Edições Sílabo, Lisboa, 1993.

[19] Nilza Nunes da Silva. *Análise probabilística: um curso introdutório; The sampling studies probabitlity: un introductory course*, volume 18. Edusp, 1998.

[20] Jeffrey S Simonoff. *Analyzing categorical data*. Springer, 2003.

[21] Wan Tang, Hua He, and Xin M Tu. *Applied categorical and count data analysis*. CRC Press, 2012.

[22] Steven K. Thompson. *Sampling*. John Wiley & Sons, Inc, 2012.

[23] Paula Vicente, Fátima Ferrão, and Elisabete Reis. *Sondagens: A amostragem como factor decisivo de qualidade*. Sílabo, Lisboa, 1996.

Anexos

Apêndice A

Inquéritos

A.1 Inquérito Particulares

ESTUDO DAS NECESSIDADES DE FORMAÇÃO NO MERCADO
Particulares

1. **Sexo**
   ☐ Masculino ☐ Feminino
2. **Idade**
   ☐ 18-35 anos ☐ 36-55 anos ☐ > 55 anos
3. **Estado civil**
   ☐ Solteiro/Divorciado/Viúvo ☐ Casado/União de facto
4. **Habilitações literárias**
   ☐ Básico
   ☐ Secundário
   ☐ Profissional
   ☐ Superior/Mestrado
   ☐ Doutoramento/Pós-graduação
5. **Situação profissional**
   ☐ Trabalhador por conta de outrém
   ☐ Trabalhador independente
   ☐ Desempregado
   ☐ Reformado
6. **Área de actividade profissional**

7. Frequenta formações usualmente ou já frequentou alguma?

☐ Sim ☐ Não

Se sim: (Se não, passar à pergunta 7.4)

7.1. Quais os motivos da formação oferecida?

☐ Incentivada pela empresa onde trabalha

☐ Actualização/aprofundamento de conhecimentos

☐ Aquisição de novos conhecimentos e competências

☐ Outros

7.2. Em que áreas?

_____

7.3. Contribuíram para o aumento do seu desempenho?

☐ Sim ☐ Não (Passar à pergunta 8)

7.3.1. Se sim, de que forma:

☐ Maior capacidade para lidar com novas tecnologias/programas informáticos

☐ Adequação/actualização dos métodos de trabalho/conhecimentos

☐ Eficácia no planeamento do trabalho

☐ Maior capacidade de delegar funções

☐ Maior capacidade de liderança

☐ Maior capacidade na gestão de conflitos

☐ Melhor dinâmica do grupo

☐ Outro _____

7.4. Quais os motivos de nunca ter frequentado formação?

☐ Nunca encontrou uma formação com interesse

☐ Elevado custo

☐ Tem de se deslocar

☐ Outros _____

7.5. Acha que poderia melhorar competências se tivesse formação?

☐ Sim ☐ Não

8. Que tipo de formação prefere?

☐ Presencial

☐ À distância (Internet)

☐ Mista

9. Qual o custo que considera justo por dia?

☐ 10-25 euros

☐ 25-55 euros

☐ 55-100 euros

☐ > 100 euros

10. Onde recorreria para obter informações sobre formações?

☐ Internet ☐ Comunicação social

☐ Publicidade ☐ Outros

11. Conhece alguma empresa que organize formações?

☐ Sim ☐ Não

12. Conhece a empresa MultiDados®/MD-Form® (serviço multidados.com na área da formação)?

☐ Sim ☐ Não

13. Estaria disponível para frequentar formações promovidas pela MD-Form®?

☐ Sim ☐ Não

14. Gostaria de nos indicar o seu endereço de email para futuras informações sobre o tema?

_____

Nome:
Local:
Telefone:
Data:

Obrigado pelo tempo dispendido!

A.2 Inquérito Empresas

ESTUDO DAS NECESSIDADES DE FORMAÇÃO NO MERCADO
Empresas

1. **Área de actividade da empresa**

_____

2. **Quantos funcionários têm por categoria profissional?**

☐ Nº funcionários em Cargos superiores (Administradores, Directores, Financeiros, ...)

☐ Nº funcionários em Cargos médios (Gestores, Chefes de Equipa, ...)

☐ Nº funcionários em Cargos técnicos (Colaboradores, Funcionários, ,...)

3. **Costuma disponibilizar formação aos funcionários?**

☐ Sim ☐ Não

Se sim: (Se não, passar à pergunta 3.6)

3.1. Com que frequência?

☐ Mensal
☐ Bimestral
☐ Semestral
☐ Anualmente

3.2. Em que áreas?

_____

3.3. Quais os motivos da formação oferecida?

☐ Actualização/aprofundamento de conhecimentos
☐ Melhoria da imagem da empresa
☐ Aumento da produtividade
☐ Recebeu subsídio específico para formação
☐ Aumento da motivação/satisfação dos colaboradores
☐ Melhoria da qualidade do serviço
☐ Diminuição do número de reclamações
☐ Outros

3.4. Considera que melhorou o desempenho dos funcionários?

☐ Sim ☐ Não

3.5. Organizam a sua própria formação ou recorrem formação externa?

☐ Formação própria ☐ Formação externa

3.6. Quais os motivos (Não disponibilizar formação aos funcionários)?

- ☐ Nunca encontrou uma formação com interesse
- ☐ Elevado custo
- ☐ Não tem instalações
- ☐ Tem de se deslocar

Se a resposta anterior foi falta de instalações:

3.6.1. Faria formação se lhe disponibilizassem instalações a um determinado custo?

☐ Sim ☐ Não

3.6.2. Qual o custo que considera justo por dia?

- ☐ 10-25 euros
- ☐ 25-50 euros
- ☐ 50-100 euros
- ☐ >100 euros

3.6.3. Que condições da sala acha necessárias?

- ☐ Quadro
- ☐ Data show
- ☐ Luz natural
- ☐ Bons acessos
- ☐ Outros

4. Importava-se de ter formação em simultâneo com outra empresa?

☐ Sim ☐ Não

5. Que tipo de formação prefere?

- ☐ Presencial
- ☐ À distância (Internet)
- ☐ Mista

6. Qual o período mais apropriado?

- ☐ Laboral
- ☐ Pós-laboral
- ☐ Sábados

7. Onde recorreria para obter informações sobre formações?

- ☐ Internet
- ☐ Publicidade
- ☐ Comunicação social
- ☐ Parcerias
- ☐ Associações comerciais
- ☐ Outros

8. Conhece alguma empresa que organize formações?

☐ Sim ☐ Não

9. Conhece a empresa MultiDados®/MD-Form® (serviço multidados.com na área da formação)?

☐ Sim ☐ Não

10. Estaria disponível para frequentar formações promovidas pela MD-Form®?

☐ Sim ☐ Não

11. Gostaria de nos indicar o seu endereço de email para futuras informações sobre o tema?

_____

Nome:

Local:

Telefone:

Data:

Obrigado pelo tempo dispendido!

Apêndice B

Informação inicial

B.1 Particulares

Concelho	Total por Região
Espinho	24252
Castelo de Paiva	13819
Arouca	19999
Oliveira de Azeméis	61170
Santa Maria da Feira	126389
São João da Madeira	18491
Vale de Cambra	21126
Águeda	42968
Albergaria-a-Velha	22565
Anadia	27290
Aveiro	61255
Estarreja	23977
Ílhavo	35602
Mealhada	19408
Murtosa	8215
Oliveira do Bairro	20271
Ovar	49446
Sever do Vouga	10844
Vagos	20827
Total	627914

População residente (N.º) por Local de residência; Decenal - INE, Censos - séries históricas

B.2 Empresas

Setor de Atividade Económica	Total por Setor
C - Indústrias transformadoras	8080
G - Comércio por grosso e a retalho; reparação de veículos automóveis e motociclos	8293
H - Transportes e armazenagem	1570
J - Actividades de informação e de comunicação	839
K - Actividades financeiras e de seguros	929
L - Actividades imobiliárias	1323
M - Actividades de consultoria, científicas, técnicas e similares	1153
Total	22187

Informação retirada do site: http://www.infoempresas.com.pt